Efficacité d'un test semi-paramétrique d'indépendance entre vecteurs aléatoires

Bernard Colin, Ernest Monga

Propriétés asymptotiques d'un test semi-paramétrique d'indépendance entre vecteurs aléatoires

Bernard Colin et Ernest Monga

Département de mathématiques
Faculté des Sciences
Université de Sherbrooke
Sherbrooke J1K-2R1 Québec Canada
bernard.colin@usherbrooke.ca
ernest.monga@usherbrooke.ca

Résumé

Étant donné n vecteurs aléatoires $X_1, X_2, ..., X_n$ de dimensions finies, on considère le test semi-paramétrique d'indépendance entre ces derniers tel que présenté dans Colin et Monga (2007, 2008). Après en avoir illustré son usage sur quelques exemples et avoir mis en évidence de façon empirique sa puissance, dans le cas de quelques hypothèses alternatives, on se propose dans un cadre plus théorique de montrer que ce test est asymptotiquement sans biais et qu'il est de plus asymptotiquement équivalent au test du rapport de vraisemblance dans le cadre d'une hypothèse alternative particulière, choix dicté par la difficulté que l'on rencontre dans ce cas précis pour décrire de façon pratique et réaliste l'hypothèse alternative.

Mots-clés : test semi-paramétrique, indépendance, vecteurs aléatoires, puissance asymptotique.

Summary

Let $X_1, X_2, ..., X_n$, be n given finite dimensional random vectors, one considers the semiparametric test of independence between these vectors as introduced by Colin and Monga (2007, 2008). After having illustrated its use on some examples and having empirically highlighted its power, in the case of some alternative hypothesis, one proposes in a more theoretical framework, to show that this test is asymptotically unbiased and that it is, moreover, asymptotically equivalent to the likelihood ratio test under some particular alternative hypothesis. This choice rises from the difficulties that one meets, in this case, to describe the alternative, in a practical and realistical way.

Key words : semiparametric test, independence, random vectors, asymptotic power.

Soient $X_i \in \mathbb{R}^{k_i}$, $i = 1, 2, ..., n$, n vecteurs aléatoires définis sur les espaces probabilisés $\left(\mathbb{R}^{k_i}, \mathcal{B}_{\mathbb{R}^{k_i}}, \mu_i, \nu_i\right)$, où $\mathcal{B}_{\mathbb{R}^{k_i}}$ μ_i et ν_i désignent respectivement la σ-algèbre de Borel de $\mathbb{R}^{k_i}$, la mesure de probabilité associée à X_i et une mesure de référence (en général la mesure de Lebesgue) telle que $\mu_i \ll \nu_i$ $\forall i = 1, 2, ..., n$, et soit l'espace probabilisé produit $\left(\times_{i=1}^{i=n}\mathbb{R}^{k_i}, \bigotimes_{i=1}^{i=n}\mathcal{B}_{\mathbb{R}^{k_i}}, \mu, \bigotimes_{i=1}^{i=n}\mu_i, \bigotimes_{i=1}^{i=n}\nu_i\right)$ où μ et $\otimes_{i=1}^{i=n}\mu_i$ désignent la mesure de probabilité conjointe de $X = (X_1, X_2, ..., X_n)$ (avec $\mu \ll \otimes_{i=1}^{i=n}\nu_i$) et la mesure de probabilité produit. On notera respectivement par $F(x_1, x_2, ..., x_n)$ et par $F_i(x_i)$ $i = 1, 2, ..., n$ les fonctions de répartition associées aux mesures μ et μ_i, $i = 1, 2, ..., n$. Enfin, pour tout $i = 1, 2, ..., n$, $O_i^-(s_i)$ désignera l'orthant négatif de $\mathbb{R}^{k_i}$ de sommet $s_i = {}^t\left(s_{i_1}, s_{i_2}, ..., s_{i_{k_i}}\right)$ c'est-à-dire, l'ensemble des points $x_i = {}^t\left(x_{i_1}, x_{i_2}, ..., x_{i_{k_i}}\right)$ de $\mathbb{R}^{k_i}$ tels que $x_{i_1} \leq s_{i_1}, x_{i_2} \leq s_{i_2}, ..., x_{i_{k_i}} \leq s_{i_{k_i}}$ que l'on notera, par commodité d'écriture par : $x_i \leq s_i$. On remarquera, de plus, que pour tout $i = 1, 2, ..., n$

$$\mu_i\left(O_i^-(s_i)\right) = F_i(s_i)$$

et que pour tout $\alpha_i \in [0, 1]$, les ensembles suivants :$\left\{s_i \in \mathbb{R}^{k_i} : \mu_i\left(O_i^-(s_i)\right) = \alpha_i\right\}$ ne sont autres que les surfaces de niveau, $F_i^{-1}(\alpha_i)$, des fonctions $F_i(x_i)$.

Dans ce qui suit, on construit un test afin de confronter les hypothèses nulle $\mathcal{H}_0$ et alternative $\mathcal{H}_1$ suivantes :

$$\mathcal{H}_0 : \mu = \bigotimes_{i=1}^{i=n}\mu_i \quad \text{et} \quad \mathcal{H}_1 : \mu \neq \bigotimes_{i=1}^{i=n}\mu_i$$

Posant $k = \sum_{i=1}^{i=n} k_i$, ces dernières, s'expriment sous la forme :

$$\mathcal{H}_0 : F(x) = \prod_{i=1}^{i=n} F_i(x_i) \quad \forall\, x \in \mathbb{R}^k \quad \text{et} \quad \mathcal{H}_1 : F(x) \neq \prod_{i=1}^{i=n} F_i(x_i)$$

où, en notant par $O^-(x)$ l'orthant négatif de $\mathbb{R}^k$ de sommet $x = {}^t(x_1, ..., x_n)$

$$\begin{aligned}
\mathcal{H}_0 \quad &: \quad \mu\left(O^-(x)\right) = \prod_{i=1}^{i=n}\mu_i\left(O_i^-(x_i)\right) \ \forall\, x \in \mathbb{R}^k \quad \text{et} \\
\mathcal{H}_1 \quad &: \quad \mu\left(O^-(x)\right) \neq \prod_{i=1}^{i=n}\mu_i\left(O_i^-(x_i)\right)
\end{aligned}$$

Soient, pour tout i, deux points s_i et t_i appartenant à $\mathbb{R}^{k_i}$ et soient $O_i^-(s_i)$ et $O_i^-(t_i)$ les orthants négatifs de sommets respectifs s_i et t_i. On dira que s_i et t_i sont équivalents si $\mu_i\left(O_i^-(s_i)\right) = \mu_i\left(O_i^-(t_i)\right) = \alpha_i$ où α_i désigne la valeur commune de ces deux probabilités. Cette égalité définit une relation d'équivalence entre les points de l'espace $\mathbb{R}^{k_i}$ et pour tout $\alpha_i \in [0, 1]$, les classes d'équivalences $\mathcal{C}_{i,\alpha_i}$, dont on désignera un représentant quelconque par s_{i,α_i}, sont les surfaces de niveau, $F_i^{-1}(\alpha_i)$, de la fonction $F_i(x_i)$. Ainsi, pour que la condition $\mu(O^-(x)) = \prod_{i=1}^{i=n}\mu_i\left(O_i^-(x_i)\right)$ soit vérifiée pour tout point $x \in \mathbb{R}^k$, il faut et il suffit qu'elle le soit pour tous les éléments d'une famille unique de

représentants s_{i,α_i} de $\mathcal{C}_{i,\alpha_i}$, où $i = 1, 2, ..., n$ et où α_i appartient à l'intervalle $[0, 1]$. La condition relative à $\mathcal{H}_0$ devient alors :

$$\begin{aligned}\mu\left(O^-\left(s_{1,\alpha_1}, s_{2,\alpha_2}, ..., s_{n,\alpha_n}\right)\right) &= \prod_{i=1}^{i=n} \mu_i\left(O_i^-\left(s_{i,\alpha_i}\right)\right) \\ &= \prod_{i=1}^{i=n} \alpha_i \quad \forall\left(\alpha_1, \alpha_2, ..., \alpha_n\right) \in [0,1]^n\end{aligned}$$

Une façon commode de choisir un représentant s_{i,α_i} par classe d'équivalence $\mathcal{C}_{i,\alpha_i}$, consiste à retenir, pour tout i et pour tout α_i, l'orthant négatif $O_i^-\left(\widetilde{s}_{i,\alpha_i}\right)$ pour lequel le sommet $\widetilde{s}_{i,\alpha_i}$ possède des coordonnées égales, dont la valeur commune sera notée par s_{α_i}. Il s'ensuit que :

$$\begin{aligned}\mu_i\left(O_i^-\left(\widetilde{s}_{i,\alpha_i}\right)\right) &= \mu_i\left(x_i \leq \widetilde{s}_{i,\alpha_i}\right) = \mathbb{P}\left(X_{i_1} \leq s_{\alpha_i}, X_{i_2} \leq s_{\alpha_i}, ..., X_{i_{k_i}} \leq s_{\alpha_i}\right) \\ &= F_{V_i}\left(s_{\alpha_i}\right)\end{aligned}$$

où $F_{V_i}(\cdot)$ désigne, pour tout i, la fonction de répartition de la variable aléatoire $V_i = Sup\left(X_{i_1}, X_{i_2}, ..., X_{i_{k_i}}\right)$. Ce choix permet de ramener le test d'indépendance d'une famille de vecteurs aléatoires au test d'indépendance des composantes d'un vecteur aléatoire puisque, dans ce cas, la condition portant sur l'hypothèse nulle $\mathcal{H}_0$ devient :

$$\mu\left(O^-\left(\widetilde{s}_{1,\alpha_1}, \widetilde{s}_{2,\alpha_2}, ..., \widetilde{s}_{n,\alpha_n}\right)\right) = \prod_{i=1}^{i=n} \mu_i O_i^-\left(\widetilde{s}_{i,\alpha_i}\right) \quad \forall\left(\alpha_1, \alpha_2, ..., \alpha_n\right) \in [0,1]^n$$

c'est-à-dire :

$$\begin{aligned}F\left(s_{\alpha_1}, s_{\alpha_2}, ..., s_{\alpha_n}\right) &= \mathbb{P}\left(X_1 \leq \widetilde{s}_{1,\alpha_1}, X_2 \leq \widetilde{s}_{2,\alpha_2}, ..., X_n \leq \widetilde{s}_{n,\alpha_n}\right) \\ &= \mathbb{P}\left(V_1 \leq s_{\alpha_1}, V_2 \leq s_{\alpha_2}, ..., V_n \leq s_{\alpha_n}\right) \\ &= \prod_{i=1}^{i=n} F_{V_i}\left(s_{\alpha_i}\right) \quad \forall\left(s_{\alpha_1}, s_{\alpha_2}, ..., s_{\alpha_n}\right) \in \mathbb{R}^n\end{aligned}$$

Posant $V = (V_1, V_2, ..., V_n)$ on sait que sous l'hypothèse nulle $\mathcal{H}_0$, la variable aléatoire $F_V(v)$ est libre et suit une loi $\mathcal{CM}_n$ à n degrés de liberté (Colin et Monga (2007)). Ce résultat permet alors de construire, dans un cadre semi-paramétrique, le test d'indépendance recherché et dont la procédure peut être décrite brièvement comme suit : on suppose que $F(V) \in \Psi(V, \theta)$ où $\Psi(V, \theta)$ désigne un famille de lois indicée par le paramètre θ. On estime alors θ par $\widehat{\theta}$ à l'aide d'un échantillon de taille m et l'on en déduit une estimation $\widehat{F}(V) \in \Psi(V, \widehat{\theta})$ de $F(V)$. On compare alors, à l'aide d'un test de type *Kolmogorov-Smirnov*, la fonction de répartition empirique $\widehat{F}(V)$ à la fonction de répartition d'une loi $\mathcal{CM}_n$. Les figures 1 et 2 ci-dessous illustrent une étude empirique de la puissance du test dans le cas de deux vecteurs normaux respectivement de $\mathbb{R}^2$ et de $\mathbb{R}^3$ et de deux vecteurs log-normaux respectivement de $\mathbb{R}^2$ et de $\mathbb{R}^3$ pour un risque α de première espèce de $0,1$ et pour une structure de dépendance donnée, dans les deux cas, par la matrice de variance-covariance Σ définie par :

$$\Sigma = \begin{bmatrix} 1 & -0,5 & r & r & r \\ -0,5 & 1 & r & r & r \\ r & r & 1 & 0,5 & 0,5 \\ r & r & 0,5 & 1 & 0,5 \\ r & r & 0,5 & 0,5 & 1 \end{bmatrix} \text{ où } |r| \leq 0,4$$

et où les tailles m d'échantillons sont $m = 20, 25, 30, 35, 40, 45, 50, 75, 100, 200$. Dans les deux cas, l'hypothèse nulle correspond à $r = 0$ et les hypothèses alternatives correspondent aux différentes valeurs de $r \neq 0$ et variant de $5/100$ en $5/100$ de $-0,4$ à $0,4$. Cette étude empirique permet de constater, tout au moins dans les cas considérés, que le test semble bien se comporter et qu'une puissance respectable est rapidement atteinte dès que $|r| \geq 0,2$ pour des échantillons de taille $m \geq 100$.

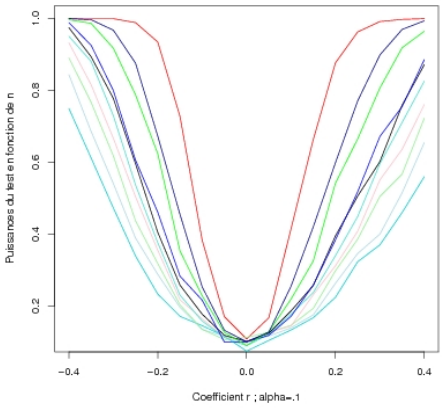

Fig.1 Vecteurs normaux

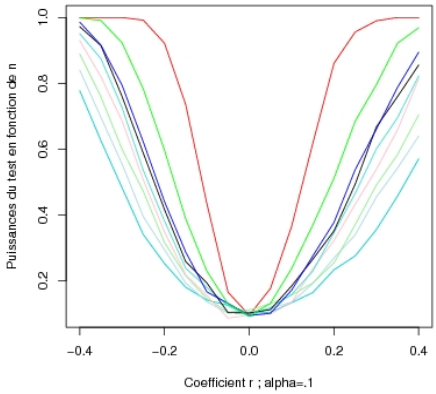

Fig.2 Vecteurs log-normaux

Quant à la figure 3, elle illustre, dans les mêmes conditions de simulation que précédemment, l'étude empirique de la puissance du test dans le cas de trois vecteurs normaux de $\mathbb{R}^2$ et pour une structure de dépendance donnée par la matrice de variance-covariance Σ suivante :

$$\Sigma = \begin{bmatrix} 1 & -0,5 & r & 0 & 0 & 0 \\ -0,5 & 1 & 0 & 0 & 0 & 0 \\ r & 0 & 1 & 0,5 & 0 & 0 \\ 0 & 0 & 0,5 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

On constate alors que les conclusions précédentes sont également valables dans ce cas.

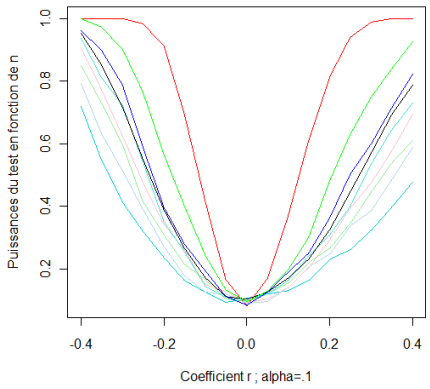

Fig.3 Cas de 3 vecteurs normaux

En ce qui concerne l'étude des propriétés asymptotiques du test, nous nous situerons dans le cadre qui suit. Soit m_ν une suite croissante d'entiers et notons par $\mathcal{H}_{0\nu}$ et $\mathcal{H}_{1\nu}$ les suites d'hypothèses nulles et alternatives associées aux m_ν observations. Par exemple, dans le cas d'observations issues de lois normales, on pourrait supposer que sous les hypothèses $\mathcal{H}_{0\nu}$ la loi des observations serait donnée par une loi $\mathcal{N}_k(0, I)$, alors que sous les hypothèses alternatives $\mathcal{H}_{1\nu}$ elles seraient décrites par des lois de la forme $\mathcal{N}_k(0, \Sigma_\nu)$. En particulier, on examinera les cas où $\lim_{\nu\to\infty} \Sigma_\nu = I$ qui est une expression de la notion de contiguïté entre les deux suites d'hypothèses. On considèrera alors les procédures de tests suivantes : la première, telle que présentée ci-dessus, pour laquelle on désignera la région critique par R_{1,m_ν} et la seconde qui correspondra au test du rapport de vraisemblance, dont la région critique sera désignée par R_{2,m_ν}. La conduite de l'étude des propriétés asymptotiques s'effectuera comme suit. Après avoir choisit deux régions critiques R_{1,m_ν} et R_{2,m_ν} telles que :

$$\mathbb{P}_{\mathcal{H}_{0\nu}}(R_{1,m_\nu}) = \mathbb{P}_{\mathcal{H}_{0\nu}}(R_{2,m_\nu}) = \alpha_\nu$$

on montrera que :

$$\lim_{\nu\to\infty} \mathbb{P}_{\mathcal{H}_{1\nu}}(R_{1,m_\nu}) = \lim_{\nu\to\infty} \mathbb{P}_{\mathcal{H}_{1\nu}}(R_{2,m_\nu}) = \beta$$

qui n'est autre que l'expression de l'équivalence asymptotique entre les deux procédures. A titre d'exemple, dans le cas de la normalité, ou plus généralement dans le cas des lois elliptiques, on considèrera des matrices de structure

de dépendance Σ_ν de la forme : $\Sigma_\nu = I_k + \frac{1}{\sqrt{m_\nu}}\Delta_\nu$ où $\lim_{\nu\to\infty} \Delta_\nu = \Delta$. Les régions critiques R_{1,m_ν} et R_{2,m_ν} associées aux procédures décrites ci-dessus seront respectivement de la forme :

$$\sup_u |\widetilde{F}_\nu(u) - G(u)| \geq c_{1,m_\nu} \quad \text{et} \quad \frac{L(\mathcal{H}_{0\nu})}{L(\mathcal{H}_{1\nu})} \leq c_{2,m_\nu}$$

où : $\widetilde{F}_\nu(u)$ désigne la fonction de répartition empirique de la variable aléatoire $F_V(v)$, $G(u)$ désigne la fonction de répartition de la loi $\mathcal{CM}_n$ à n degrés de liberté et où $\frac{L(\mathcal{H}_{0\nu})}{L(\mathcal{H}_{1\nu})}$ désigne le rapport de vraisemblance généralisé habituel.

Bibliographie

[1] Bernard Colin et Ernest Monga (2007) : *Test semi-paramétrique d'indépendance des composantes d'un vecteur aléatoire.* Pub.Inst.Stat.Univ. Paris LI fasc.1-2 3 à 24.

[2] Bernard Colin et Ernest Monga (2008) : *Semiparametric Test of Indepedence between Random Vectors* Proceedings of the SFdS-SSC joint meeting. Ottawa.