Aspect brownien d'un test semi-paramétrique d'indépendance

Bernard Colin, Ernest Monga

▶ **To cite this version:**

Bernard Colin, Ernest Monga. Aspect brownien d'un test semi-paramétrique d'indépendance. 42èmes Journées de Statistique, 2010, Marseille, France, France. inria-00494775

HAL Id: inria-00494775
https://inria.hal.science/inria-00494775

Submitted on 24 Jun 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Aspect brownien d'un test semi-paramétrique d'indépendance

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 (2009). Après en avoir illustré son usage sur quelques exemples et avoir mis en évidence de façon empirique sa puissance, on se propose alors dans un cadre plus théorique, de montrer que ce test est naturellement associé à un pont brownien, ce qui permet ainsi, dans le cas de certaines formes d'hypothèses alternatives de décrire plus aisément son comportement asymptotique. Enfin, on comparera au niveau de l'efficacité relative, le test proposé au test du rapport de vraisemblance dans le cadre de certains modèles paramétriques donnés.

Mots-clés : test semi-paramétrique, indépendance, vecteurs aléatoires, pont brownien, 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 (2009). 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 naturally linked with a brownian bridge which allow an easy description of its asymptotic behaviour for some alternatives. Finally the proposed test will be compared to some competitors using relative efficiency.

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

Généralités

Soient $X_i \in \mathbb{R}^{k_i}$, $i = 1, 2, ..., n$, n vecteurs aléatoires définis sur les espaces probabilisés $\left(\mathbb{R}^{k_i}, \mathcal{B}_{\mathbb{R}^{k_i}}, \mu_i \ll \nu_i\right)$ où μ_i désigne la mesure de probabilité associée à X_i et ν_i désigne une mesure de référence. Si l'on note par μ la mesure de probabilité conjointe des vecteurs $(X_1, X_2, ..., X_n)$, définie sur $\bigotimes_{i=1}^{i=n} \mathcal{B}_{\mathbb{R}^{k_i}}$, on considère alors les hypothèses nulle et alternatives suivantes :

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

Utilisant la notion d'orthant négatif, Colin et Monga (2007) ont montré que le test d'indépendance entre les vecteurs aléatoires, se ramenait à un test d'indépendance entre les composantes d'un vecteur $V =$

$(V_1, V_2, ..., V_n)$ de fonction de répartition $F(V)$ dont la loi sous l'hypothèse nulle $\mathcal{H}_0$ est libre et admet pour fonction de répartition $G_n(z)$ l'expression définie par :

$$G_n(z) = z \sum_{i=0}^{n-1} \frac{[Log(1/z)]^i}{i!} \mathbb{I}_{[0,1]}(z) + \mathbb{I}_{]1,\infty[}(z)$$

Si l'on suppose que $F(V) \in \Psi(V, \theta)$ où $\Psi(V, \theta)$ désigne une famille de lois indicées par le paramètre θ et si $\hat{\theta}$ est un estimateur de θ on peut en déduire une estimation $\widehat{F}(V) \in \Psi(V, \hat{\theta})$ 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 ci-dessus. Les figures 1 et 2 suivantes 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 :

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

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 $r \neq 0$ 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 $m \geq 100$.

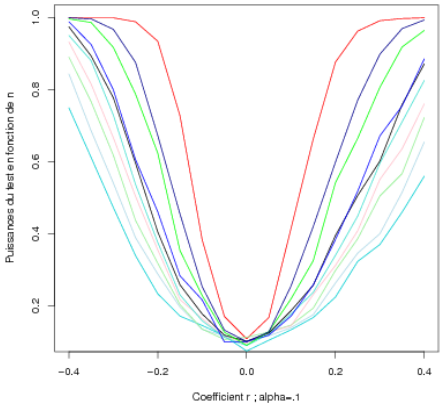

Fig.1 Vecteurs normaux

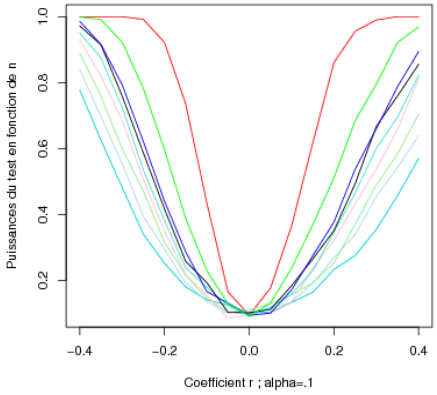

Fig.2 Vecteurs log-normaux

Test et pont brownien

Puisque le test d'indépendance entre vecteurs aléatoires se réduit à celui d'un test d'indépendance entre les composantes d'un vecteur aléatoire, il suffit, aux fins de la présente étude, de se restreindre à ce dernier cas. On supposera pour la suite que l'on dispose de n observations indépendantes du vecteur aléatoire $X = (X_1, X_2, ..., X_k)$. Désignant par $F_U(u)$ la fonction de répartition inconnue de la variable aléatoire $U = G_k(F(X_1, X_2, ..., X_k))$ et par $\hat{F}_{U,n}(u)$ sa version empirique, la procédure de test revient à considérer la quantité Δ_n définie par :

$$\Delta_n = \sup_{0 \leq u \leq 1} |\hat{F}_{U,n}(u) - u|$$

et à rejeter l'hypothèse d'indépendance pour les valeurs de Δ_n excédant un certain seuil δ_α. En d'autres termes, le test proposé peut, à l'aide de la variable aléatoire $U = (G_k \circ F)(X)$ à valeurs dans $[0,1]$, se formuler comme suit :

$$\begin{aligned} \mathcal{H}_0 &: U \text{ suit une loi uniforme } \mathcal{U}_{[0,1]} \\ \mathcal{H}_1 &: U \text{ ne suit pas une loi uniforme } \mathcal{U}_{[0,1]} \end{aligned}$$

et se ramène donc à un test d'adéquation à la loi $\mathcal{U}_{[0,1]}$ de type *Kolmogorov-Smirnov*.

Posant alors, pour tout $0 \leq u \leq 1$:

$$Y_n(u) = \sqrt{n}(\hat{F}_{U,n}(u) - u) \ ,$$

il vient :

$$\sqrt{n}\Delta_n = \sqrt{n} \sup_{0 \leq u \leq 1} |\hat{F}_{U,n}(u) - u| = \sup_{0 \leq u \leq 1} |Y_n(u)| \ .$$

Le test proposé s'exprime donc naturellement sous la forme d'un "pont brownien" $W^0(u)$ pour lequel, sous l'hypothèse nulle $\mathcal{H}_0$, on a (voir Billingsley (1968), Le Cam (1986)) :

$$\sqrt{n}\Delta_n \xrightarrow{d} \sup_{0 \leq u \leq 1} |W^0(u)| \ ,$$

avec :

$$\begin{aligned} \lim_{n \to \infty} \mathbb{P}(\sqrt{n}\Delta_n \quad &\leq \quad d) = W^0\{x(\cdot) : \sup_{0 \leq u \leq 1} |x(u)| \leq d\} \ , \\ &= \quad 1 - 2\textstyle\sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2 d^2} \quad d > 0 \ . \end{aligned}$$

Il s'ensuit alors que le comportement asymptotique du test considéré dans le cas présent, découlera directement des propriétés similaires d'un pont brownien.

Puissance du test

Dans un contexte où l'hypothèse alternative revêt une forme complexe, l'étude de la puissance du test d'indépendance s'avère ardue, dans la mesure où la "non-indépendance" qui caractérise l'hypothèse alternative, ne peut en aucun cas se décrire systématiquement. Dans ce qui suit, on s'intéresse à une suite d'hypothèses alternatives simples se ramenant, selon le comportement de la quantité :

$$d_K(F, u) = \sup_{0 \leq u \leq 1} |F(u) - u|$$

où plus généralement selon le comportement de la quantité :

$$d_K(F, F_0) = \sup_{0 \leq u \leq 1} |F(u) - F_0(u)|$$

(où F_0 correspond à une hypothèse nulle $\mathcal{H}_0 : F = F_0$ quelconque), aux trois cas présentés ci-après. Les résultats rappelés ci-dessous, et dont découlent les propriétés asymptotiques du test proposé, se trouvent exposés en particulier, dans : Chibisov (1965), Lehmann (2005) et Massey (1950) et concernent, dans chacun des cas, la puissance asymptotique du test de *Kolmogorov-Smirnov*, pour une famille d'alternatives particulières.

- **1^{er} cas :** $\lim_{n \to \infty} n^{1/2} d_K(F_n, F_0) = \infty$.

On considère une suite $X_1, X_2, ..., X_n$ de variables aléatoires indépendantes et de même loi, de fonction de répartition F. On montre alors que le test de *Kolmogorov-Smirnov* pour l'hypothèse nulle $\mathcal{H}_0 : F = F_0$ contre l'hypothèse alternative $\mathcal{H}_1 : F \neq F_0$, possède une puissance qui tend uniformément vers 1 sur le sous-ensemble de toutes les alternatives F telles que :

$$n^{1/2} d_K(F, F_0) \geq \delta_n \ ,$$

où δ_n tend vers l'infini avec n. En d'autres termes,

$$\inf\{\mathbb{P}_F(n^{1/2}\Delta_n > d_{n,\alpha}) : n^{1/2} d_K(F, F_0) \geq \delta_n\}$$

tend vers 1 si $\delta_n \to \infty$ et où $d_{n,\alpha}$ représente la valeur critique pour un risque α de première espèce fixé et pour une taille d'échantillon n donnée.

Cette condition n'empêche nullement qu'une suite $\{F_n\}_{n\geq 1}$ d'alternatives, puisse converger au sens de la métrique de la convergence uniforme, vers l'hypothèse nulle F_0, à condition toutefois que cette dernière se fasse à une vitesse contrôlée par δ_n. Par exemple, si δ_n est de la forme n^ϵ avec $\epsilon > 0$, alors la distance entre F_n et F_0 devra être, au moins, de l'ordre de $n^{-1/2+\epsilon}$.

Soit alors $\{F_n\}_{n\geq 1}$ une suite quelconque d'hypothèses alternatives satisfaisant à la condition

$$n^{1/2} d_K(F_n, F_0) \geq \delta_n \ .$$

En vertu de l'inégalité du triangle, il vient :

$$d_K(F_n, F_0) \leq d_K(\hat{F}_n, F_0) + d_K(F_n, \hat{F}_n) \ ,$$

où $\hat{F}_n$ désigne la fonction de répartiton empirique de F_n. De cette dernière inégalité on en déduit que ;

$$n^{1/2}\Delta_n \geq \delta_n - n^{1/2} d_K(F_n, \hat{F}_n) \ .$$

Il s'ensuit alors que :

$$1 \geq \mathbb{P}_{F_n}(n^{1/2}\Delta_n > d_{n,\alpha}) \geq \mathbb{P}_{F_n}(n^{1/2} d_K(F_n, \hat{F}_n) \leq \delta_n - d_{n,\alpha})$$

Or, sous l'hypothèse F_n, on montre que (voir Lehmann (2005)) la suite des variables aléatoires $n^{1/2} d_K(F_n, \hat{F}_n)$ est tendue (on dit aussi bornée en probabilité, ce que l'on note par : $n^{1/2} d_K(F_n, \hat{F}_n) = O_{F_n}(1)$). De plus :

$$\lim_{n\to\theta} d_{n,\alpha} = d_\alpha < \infty \ ,$$

où d_α correspond à la frontière de la région critique pour la loi limite de $n^{1/2}\Delta_n$ sous l'hypothèsse nulle. Comme δ_n tend vers l'infini avec n, alors $\delta_n - d_{n,\alpha} \to \infty$ et

$$\lim_{n\to\infty} \mathbb{P}_{F_n}(n^{1/2}\Delta_n > d_{n,\alpha}) = 1 \ .$$

A distance finie, c'est-à-dire lorsque l'on fixe la valeur de n, on ne dispose pas de la valeur exacte de la puissance, mais il est possible d'obtenir une borne inférieure de celle-ci à l'aide de l'inégalité de *Dvoretsky-Kiefer-Wolfowitz* (voir par exemple Bosq et Lecoutre (1987)) donnée par :

$$\forall d > 0 : \mathbb{P}_F\{d_K(\hat{F}_n, F) > d\} \leq Ce^{-2nd^2} \ ,$$

où C est une constante universelle, que l'on peut choisir égale à 2 (Massart (1990)). On montre alors que l'on obtient à l'aide de cette inégalité, une borne inférieure pour la puissance du test et pour l'alternative F_n fixée donnée par ;

$$\mathbb{P}_{F_n}(n^{1/2}\Delta_n > d_{n,\alpha}) \geq 1 - 2e^{-2(\delta_n - d_{n,\alpha})^2}$$

à condition que l'on ait : $n^{1/2} d_K(F_n, F_0) \geq \delta_n$ et $\delta_n > d_{n,\alpha}$.

- **$2^{ème}$ cas :** $\lim_{n\to\infty} n^{1/2} d_K(F_n, F_0) = 0.$

Le résultat ci-dessus montre que la puissance asymptotique du test de *Kolmogorov* et, par voie de conséquence, la puissance asymptotique du test d'indépendance, est abitrairement proche de 1 pour toute suite d'alternatives $\{F_n\}_{n\geq 1}$ tendant vers $F_0 = I_d$ ($F_0(u) = I_d(u) = u : 0 \leq u \leq 1$) suffisamment lentement. A l'opposé, l'inégalité du triangle permet d'écrire que :

$$d_K(\hat{F}_n, F_0) \leq d_K(\hat{F}_n, F) + d_K(F, F_0)$$

d'où :

$$n^{1/2}\Delta_n \leq n^{1/2} d_K(\hat{F}_n, F) + n^{1/2} d_K(F, F_0)$$

soit encore :

$$\mathbb{P}_F(n^{1/2}\Delta_n > d_{n,\alpha}) \leq \mathbb{P}_F\{n^{1/2} d_K(\hat{F}_n, F) + n^{1/2} d_K(F, F_0) > d_{n,\alpha}\}$$

ce qui entraîne une puissance médiocre du test dans le cas d'une convergence trop rapide de la suite d'alternatives $\{F_n\}_{n\geq 1}$ vers F_0 ($F_0(u) = u$, dans le cas du test d'indépendance). Plus précisement on a le résultat suivant (voir Lehmann (2005)) :

Le test de *Kolmogorov-Smirnov*, de niveau α, pour l'hypothèse $\mathcal{H}_0 : F = F_0$ contre l'hypothèse $\mathcal{H}_1 : F \neq F_0$, possède une puissance asymptotique inférieure ou égale à α pour toute suite $\{F_n\}_{n\geq 1}$ d'hypothèses alternatives telles que :

$$\lim_{n\to\infty} n^{1/2} d_K(F_n, F_0) = 0$$

ce qui, en d'autres termes, peut s'écrire :

$$\limsup_n \mathbb{P}_{F_n}\{n^{1/2}\Delta_n > d_{n,\alpha}\} \leq \alpha$$

Par conséquent le test d'indépendance ne peut distinguer des suites de contre-hypothèses qui se trouveraient à une distance de l'ordre de $o(n^{-1/2})$ de F_0.

- **$3^{ème}$ cas :** $\lim_{n\to\infty} n^{1/2} d_K(F_n, F_0) = \delta$, où $0 < \delta < \infty$.

On suppose qu'il existe de plus (voir Chibisov (1965), Lehmann (2005)), une fonction d telle que :

$$\sup_u |d_n(u) - d(u)| \longrightarrow 0 \ , \text{ où } d_n(u) = n^{1/2}\left[F_n(u) - F_0(u)\right]$$

Il vient :

$$n^{1/2}\left[\hat{F}_n(u) - F_0(u)\right] = n^{1/2}\left[\hat{F}_n(u) - F_n(u)\right] + d_n(u) \ .$$

Sous l'hypothèse alternative $F = F_n$, $n^{1/2}\left[\hat{F}_n(u) - F_n(u)\right]$ admet une moyenne nulle et une variance donnée par :

$$F_n(u)\left[1 - F_n(u)\right]$$

laquelle converge vers $F_0(u)\left[1 - F_0(u)\right]$ lorsque $n \longrightarrow \infty$.

Par ailleurs, pour tout u fixé, le théorème central limite entraîne que, sous l'hypothèse $F = F_n$:

$$n^{1/2}\left[\hat{F}_n(u) - F_n(u)\right] \xrightarrow{d} B(u) \ ,$$

où $B(u)$ est un pont brownien de loi normale $\mathcal{N}(0, F_0(u)\left[1 - F_0(u)\right])$. Ainsi sous l'hypothèse alternative,

$$n^{1/2}\left[\hat{F}_n(u) - F_0(u)\right] \xrightarrow{d} B(u) + d(u) \sim \mathcal{N}(d(u), F_0(u)\left[1 - F_0(u)\right]) \ .$$

On en déduit alors aisément que, sous l'hypothèse $F = F_n$ et pour tout $u_1, u_2, ..., u_k$, on a :

$$n^{1/2}\left[\hat{F}_n(u_1) - F_0(u_1), ..., \hat{F}_n(u_k) - F_0(u_k)\right] \xrightarrow{d} [B(u_1) + d(u_1), ..., B(u_k) + d(u_k)] \ ,$$

d'où il découle que :

$$\max_{i=1,2,...,k} n^{1/2}|\hat{F}_n(u_i) - F_0(u_i)| \xrightarrow{d} \max_{i-1,2,...,k} |B(u_i) + d(u_i)| \ ,$$

ce qui entraîne, toujours sous l'hypothèse $F = F_n$, que :

$$\sup_u n^{1/2}|\hat{F}_n(u) - F_0(u)| \xrightarrow{d} \sup_u |B(u) + d(u)|$$

Il est alors possible d'en déduire que la puissance du test d'indépendance tend vers 1 lorsque n tend vers l'infini.

Bibliographie

[1] B. Colin et E. Monga (2009) : *Efficacité d'un test semi-paramétrique d'indépendance entre vecteurs aléatoires*, Congrès de la SFdS : $41^{èmes}$ Journées de Statistique (Bordeaux).

[2] B. Colin et E. 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.

[3] P. Billingsley, (1968). *Convergence of Probability Measures*, John Wiley.

[4] L. Le Cam, (1986). *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag.

[5] D.M. Chibisov, (1965). *An investigation of the asymptotic power of the tests of fit*, Th. Prob. Applic., **10** 421-437.

[6] E.L. Lehmann and J.P. Romano, (2005). *Testing Statistical Hypotheses*, Third Edition, Springer.

[7] F.J. Massey, (1950). *A note on the power of a non-parametric test*, Annals of Mathematical Statistics, **21**, 440-443.

[8] D. Bosq et J.P. Lecoutre, (1987). *Théorie de l'estimation fonctionnelle,* Economica.

[9] P. Massart, (1990). *The tight constant in the Dvoretsky-Kiefer-Wolfowitz inequality*, Annals of Probability,
18, 1269-1283.