Tests d'uniformité sur la sphère unité de grande dimension

Davy Paindaveine[1], Christine Cutting[2] & Thomas Verdebout[3]

[1] *Université libre de Bruxelles, ECARES and Département de Mathématique, Avenue Roosevelt, 50, ECARES, CP114/04, B-1050, Bruxelles, Belgique, dpaindav@ulb.ac.be*
[2] *Université libre de Bruxelles, Département de Mathématique, Boulevard du Triomphe, Campus de la Plaine, CP210, B-1050, Bruxelles, Belgique, Christine.Cutting@ulb.ac.be*
[3] *Université libre de Bruxelles, Département de Mathématique, Boulevard du Triomphe, Campus de la Plaine, CP210, B-1050, Bruxelles, Belgique, tverdebo@ulb.ac.be*

Résumé. Nous considérons le problème de test d'uniformité sur la sphère unité en grande dimension. Notre intérêt se porte principalement sur les propriétés de puissance. A cette fin, nous considérons des contre-hypothèses à symétrie rotationnelle et nous identifions les hypothèses contiguës à l'hypothèse nulle d'uniformité. Ceci révèle une structure de normalité locale et asymptotique (LAN), qui, pour la première fois, permet de recourir au troisième lemme de Le Cam en grande dimension. Sous des conditions très faibles, nous obtenons la loi asymptotique non nulle du test de Rayleigh en grande dimension et montrons que ce test mène à des taux de convergence plus lents. Tous nos résultats (n, p)-asymptotiques sont "universels", dans le sens que la dimension p peut aller vers l'infini de façon arbitraire en fonction de la taille d'échantillon n. Une partie de nos résultats couvre également le cas de petite dimension, ce qui permet d'expliquer heuristiquement le comportement asymptotique du test de Rayleigh en grande dimension. Une étude de Monte Carlo confirme nos résultats asymptotiques.

Mots-clés. Contiguïté, grande dimension, normalité locale et asymptotique, loi à symétrie rotationnelle, statistique directionnelle, tests d'uniformité

Abstract. We consider the problem of testing for uniformity on high-dimensional unit spheres. We are primarily interested on non-null issues. To this end, we consider rotationally symmetric alternatives and identify alternatives that are contiguous to the null of uniformity. This reveals a Locally and Asymptotically Normality (LAN) structure, which, for the first time, allows to use Le Cam's Third lemma in the high-dimensional setup. Under very mild assumptions, we derive the asymptotic non-null distribution of the high-dimensional Rayleigh test and show that this test actually exhibits slower consistency rates. All (n, p)-asymptotic results we derive are "universal", in the sense that the dimension p is allowed to go to infinity in an arbitrary way as a function of the sample size n. Part of our results also cover the low-dimensional case, which also allows to explain heuristically the high-dimensional non-null behavior of the Rayleigh test. A Monte Carlo study confirms our asymptotic results.

Keywords. Contiguity, high-dimensional statistics, local and asymptotic normality, rotationally symmetric distributions, spherical statistics, tests of uniformity

1 Introduction

Lorsque des observations $\mathbf{X}_{in}$, $i = 1, \ldots, n$ à valeurs sur la sphère unité $\mathcal{S}^{p_n-1} := \{\mathbf{x} \in \mathbb{R}^{p_n} : \|\mathbf{x}\| = \sqrt{\mathbf{x}'\mathbf{x}} = 1\}$ sont disponibles, le test d'uniformité sur $\mathcal{S}^{p_n-1}$ le plus célèbre est le test de Rayleigh (1919), qui rejette l'hypothèse nulle $\mathcal{H}_{0n}$ pour de grandes valeurs de

$$R_n := np_n\|\bar{\mathbf{X}}_n\|^2 = p_n + \frac{2p_n}{n}\sum_{1\leq i<j\leq n}\mathbf{X}'_{ni}\mathbf{X}_{nj},$$

où $\bar{\mathbf{X}}_n := \frac{1}{n}\sum_{i=1}^n \mathbf{X}_{ni}$. Si $p_n = p$ pour tout n, le test est basé sur la loi χ^2_p que suit asymptotiquement R_n sous $\mathcal{H}_{0n}$. Dans le cas à grande dimension (pour lequel $p_n \to \infty$ lorsque $n \to \infty$), Paindaveine et Verdebout (2015) ont montré que, après standardisation, la statistique de Rayleigh R_n est asymptotiquement normale standard sous l'hypothèse nulle. Plus précisément, ils ont établi le résultat suivant.

Théorème 1 *Soit (p_n) une suite d'entiers positifs telle que $p_n \to \infty$ si $n \to \infty$. Supposons que $\mathbf{X}_{ni}$, $i = 1, \ldots, n$, $n = 1, 2, \ldots$, soit un array triangulaire de vecteurs aléatoires tels que, pour chaque n, $\mathbf{X}_{n1}, \mathbf{X}_{n2}, \ldots, \mathbf{X}_{nn}$ soient mutuellement indépendants et soient tous uniformément distribués sur $\mathcal{S}^{p_n-1}$. Alors*

$$R^{\mathrm{St}}_n := \frac{R_n - p_n}{\sqrt{2p_n}} = \frac{\sqrt{2p_n}}{n}\sum_{1\leq i<j\leq n}\mathbf{X}'_{ni}\mathbf{X}_{nj} \tag{1}$$

converge faiblement vers la loi normale standard lorsque $n \to \infty$.

Le test de Rayleigh en grande dimension qui résulte de ce théorème consiste à rejeter $\mathcal{H}_{0n}$ au niveau asymptotique α si R^{St}_n excède le quantile d'ordre $1-\alpha$ de la loi normale standard. Ce test est valide de façon "universelle", dans le sens où sa dimension asymptotique sous $\mathcal{H}_{0n}$ vaut α quelle que soit la façon dont p_n tend vers l'infini avec n. Ce test peut donc être utilisé dès que p_n et n sont grands, sans se soucier de leur magnitude relative. Ceci est en contraste avec la plupart des résultats asymptotiques en grande dimension, qui demandent typiquement que $p_n/n \to c$, pour un certain c dans $(0, \infty)$. De plus, une loi asymptotique commune est obtenue pour tous les (n, p)-régimes, au contraire, par exemple, des tests proposés par Cai et Jiang (2012) ou Cai, Fan et Jiang (2013).

Ces bonnes propriétés sous l'hypothèse ne permettent bien entendu pas à elles seules de justifier le recours au test de Rayleigh. En effet, le test trivial, qui rejette l'hypothèse lorsqu'une variable aléatoire uniforme sur $(0, 1)$ et indépendante des observations prend une valeur inférieure à α, est aussi de dimension α sous l'hypothèse, et ce de façon universelle, mais ce test ne montre de puissance sous aucune contre-hypothèse. Pour établir que le test de Rayleigh est une procédure satisfaisante, il convient donc d'étudier son risque de seconde espèce, ce qui est l'un des objectifs principaux de ce travail.

2 Contribution

Nous considérons des contre-hypothèses à symétrie rotationnelle. La loi d'un vecteur aléatoire $\mathbf{X}$ à valeurs sur $\mathcal{S}^{p-1}$ est dite être à symétrie rotationnelle par rapport à $\boldsymbol{\theta}(\in \mathcal{S}^{p-1})$ si et seulement si $\mathbf{OX}$ a la même distribution que $\mathbf{X}$ pour toute matrice $p\times p$ orthogonale $\mathbf{O}$ vérifiant $\mathbf{O}\boldsymbol{\theta}=\boldsymbol{\theta}$; voir Saw (1978). Une telle loi, qui est complètement caractérisée par le centre de symétrie $\boldsymbol{\theta}$ et la fonction de répartition F de $\mathbf{X}'\boldsymbol{\theta}$, sera désignée par $\mathcal{R}_p(\boldsymbol{\theta}, F)$.

Plus spécifiquement, nous considérons des arrays triangulaires de la forme $\mathbf{X}_{ni}$, $i = 1,\ldots,n$, $n = 1,2,\ldots$ tels que, pour tout n, les vecteurs aléatoires $\mathbf{X}_{n1}, \mathbf{X}_{n2}, \ldots, \mathbf{X}_{nn}$ sont indépendants et identiquement distribués de loi $\mathcal{R}_{p_n}(\boldsymbol{\theta}_n, F_n)$, pour une certaine suite $(\boldsymbol{\theta}_n)$ de vecteurs de $\mathcal{S}^{p_n-1}$ et de fonctions de répartition (F_n) sur $[-1,1]$. L'hypothèse correspondante sera désignée par $\mathrm{P}^{(n)}_{\boldsymbol{\theta}_n,F_n}$. Dans la suite, nous noterons respectivement $e_{n\ell} := \mathrm{E}[(\mathbf{X}'_{ni}\boldsymbol{\theta}_n)^\ell]$ et $\tilde{e}_{n\ell} := \mathrm{E}[(\mathbf{X}'_{ni}\boldsymbol{\theta}_n - e_{n1})^\ell]$ les moments d'ordre ℓ non centrés et centrés associés à F_n et nous poserons $f_{n\ell} := \mathrm{E}[(1-(\mathbf{X}'_{ni}\boldsymbol{\theta}_n)^2)^{\ell/2}]$; toutes les espérances sont ici prises sous $\mathrm{P}^{(n)}_{\boldsymbol{\theta}_n,F_n}$.

Le résultat suivant précise la loi asymptotique du test de Rayleigh en grande dimension sous les contre-hypothèses ci-dessus.

Théorème 2 *Soit (p_n) une suite d'entiers positifs telle que $p_n\to\infty$ si $n\to\infty$. Soient $(\boldsymbol{\theta}_n)$ une suite telle que $\boldsymbol{\theta}_n\in\mathcal{S}^{p_n-1}$ pour tout n et F_n une suite de fonctions de répartition sur $[-1,1]$ telle que*

$$(i)\ \min\Big(\frac{p_n\tilde{e}^2_{n2}}{f^2_{n2}}, \frac{\tilde{e}_{n2}}{ne^2_{n1}}\Big) = o(1), \quad (ii)\ \tilde{e}_{n4}/\tilde{e}^2_{n2} = o(n), \quad \text{and} \quad (iii)\ f_{n4}/f^2_{n2} = o(n)$$

pour $n\to\infty$. Alors, en posant $\sigma^2_n := p_n\tilde{e}^2_{n2} + 2np_ne^2_{n1}\tilde{e}_{n2} + f^2_{n2}$, on a que, sous $\mathrm{P}^{(n)}_{\boldsymbol{\theta}_n,F_n}$,

$$\frac{R^{\mathrm{St}}_n - \mathrm{E}[R^{\mathrm{St}}_n]}{\sigma_n} = \frac{\sqrt{2p_n}}{n\sigma_n}\sum_{1\le i<j\le n}\big(\mathbf{X}'_{ni}\mathbf{X}_{nj} - e^2_{n1}\big)$$

converge faiblement vers la loi normale standard lorsque $n\to\infty$.

Ce résultat permet d'étudier de façon assez fine la puissance du test de Rayleigh en grande dimension. Nous ne présentons ici que le corollaire suivant, pour la bonne lecture duquel nous indiquons que, sous $\mathcal{H}_{0n}$, nous avons $e_{n1}=0$ et $\tilde{e}_{n2}=1/p_n$.

Corollaire 1 *Soit (p_n) une suite d'entiers positifs telle que $p_n\to\infty$ si $n\to\infty$. Soient $(\boldsymbol{\theta}_n)$ une suite telle que $\boldsymbol{\theta}_n\in\mathcal{S}^{p_n-1}$ pour tout n et F_n une suite de fonctions de répartition sur $[-1,1]$ satisfaisant les hypothèses du Théorème 2 et telle que, pour un certain réel τ,*

$$e_{n1} = \frac{\tau}{n^{1/2}p_n^{1/4}} + o\Big(\frac{1}{n^{1/2}p_n^{1/4}}\Big) \qquad \text{et} \qquad \tilde{e}_{n2} = \frac{1}{p_n} + o\Big(\frac{1}{p_n}\Big).$$

Alors, sous $\mathrm{P}^{(n)}_{\boldsymbol{\theta}_n,F_n}$, la puissance asymptotique du test de Rayleigh est donnée par $1-\Phi\big(\Phi(1-\alpha)-\tau^2/\sqrt{2}\big)$, où Φ désigne la fonction de répartition de la loi normale standard.

Notre travail ne s'arrête pas identifier ces contre-hypothèses sous lesquelles le test de Rayleigh montre des puissances asymptotiques non triviales en grande dimension, mais étudie surtout la suite d'expériences statistiques sous-jacentes en termes de contiguïté et de normalité locale et asymptotique.

Bibliographie

[1] Cai, T., Fan, J., et Jiang, T. (2013), Distributions of angles in random packing on spheres, *J. Mach. Learn. Res.*, 14, 1837–1864.
[2] Cai, T., et Jiang, T. (2012), Phase transition in limiting distributions of coherence of high-dimensional random matrices, *J. Multivariate Anal.*, 107, 24–39.
[3] Paindaveine, D. et Verdebout, Th. (2015), Universal asymptotics for high-dimensional sign tests. Soumis.
[4] Rayleigh, L. (1919), On the problem of random vibrations and random flights in one, two and three dimensions, *Phil. Mag.*, 37, 321–346.
[5] Saw, J.G. (1978), A family of distributions on the m-sphere and some hypothesis tests, Biometrika, 65, 69–73.