Tests d'indépendance et tests d'homogénéité basés sur des méthodes à noyaux

Mélisande Albert [1], Béatrice Laurent [1], Amandine Marrel [2], Anouar Meynaoui [3], Antonin Schrab [4], Ilmun Kim [5], Benjamin Guedj [4] & Arthur Gretton [6]

[1] *Institut de Mathématiques de Toulouse, INSA Toulouse, melisande.albert@insa-toulouse.fr, beatrice.laurent@insa-toulouse.fr*
[2] *CEA, DES, IRESNE, DER, SESI, Cadarache, associée à l'IMT, amandine.marrel@cea.fr*
[3] *Institut de Recherche Mathématique de Rennes, Université de Rennes 2, anouar.meynaoui@univ-rennes2.fr*
[4] *Centre for Artificial Intelligence, University College London & Inria London, a.schrab@ucl.ac.uk, b.guedj@ucl.ac.uk*
[5] *Department of Statistics & Data Science, Yonsei University, ilmun@yonsei.ac.kr*
[6] *Gatsby Computational Neuroscience Unit, University College London, arthur.gretton@gmail.com*

Résumé. Cet exposé s'appuie sur deux articles respectivement en collaboration avec M. Albert, A. Marrel et A. Meynaoui [Albert et al., 2022] et avec A. Schrab, I. Kim, M. Albert, B. Guedj et A. Gretton [Schrab et al., 2023]. Nous nous intéressons d'une part à tester l'indépendance de deux vecteurs $X \in \mathbb{R}^p$ et $Y \in \mathbb{R}^q$ à partir de l'observation d'un n-échantillon $((X_1, Y_1), \dots, (X_n, Y_n))$ et d'autre part à tester que deux échantillons indépendants de variables aléatoires à valeurs dans $\mathbb{R}^p$, $(X_1, \dots, X_m)$ i.i.d. de loi de probabilité P et $(Y_1, \dots, Y_n)$ i.i.d. de loi de probabilité Q, ont même loi. Le point commun de ces deux papiers est d'utiliser la notion de MMD (Maximum Mean Discrepancy) qui définit une métrique entre lois de probabilités basée sur des noyaux dans des espaces de Hilbert à noyau reproduisant (RKHS). Plus précisément, étant donné un RKHS $\mathcal{H}_k$ associé au noyau k, la MMD entre deux mesures de probabilités P et Q est définie par

$$\mathrm{MMD}(P, Q, \mathcal{H}_k) = \sup_{f \in \mathcal{H}_k, \|f\|_{\mathcal{H}_k} \leq 1} \mathbb{E}_{X \sim P}\left[f(X)\right] - \mathbb{E}_{Y \sim Q}\left[f(Y)\right].$$

Pour certains types de noyaux, dits caractéristiques, la nullité de $\mathrm{MMD}(P, Q, \mathcal{H}_k)$ équivaut à l'égalité des mesures de probabilités P et Q.

Pour le problème de test d'égalité des lois P et Q de deux échantillons, nous nous concentrons sur l'estimation de la quantité $\mathrm{MMD}(P, Q, \mathcal{H}_k)$, pour un certain choix de noyau k, en suivant les travaux précurseurs de [Gretton et al., 2007]. Par ailleurs, pour tester l'indépendance de deux vecteurs aléatoires X et Y, nous proposons d'utiliser le critère d'indépendance de Hilbert-Schmidt (HSIC) qui a été introduit par [Gretton et al., 2005]), et qui n'est autre que la MMD (associée à un certain noyau) entre la loi du couple (X, Y) et le produit des lois marginales.

L'objectif de l'exposé sera de montrer comment on peut construire des estimateurs de la MMD et du HSIC puis d'en déduire des tests d'homogénéité et des tests d'indépendance.

Nous verrons en particulier le recours à des techniques de permutation pour garantir le niveau des tests. Par ailleurs, nous donnerons des résultats de puissance pour les tests, qui s'appuient sur des inégalités exponentielles pour les U-statistiques dues à [Arcones and Giné, 1993] et [Giné et al., 2000]. Nous discuterons également du choix des noyaux utilisés et nous montrerons l'intérêt d'agréger des tests associés à différents noyaux.

Mots-clés. Tests non paramétriques, Maximum Mean Discrepancy, critère d'indépendance de Hilbert-Schmidt, méthodes de permutation, tests agrégés.

Abstract. This presentation is based on two papers respectively in collaboration with M. Albert, A. Marrel and A. Meynaoui [Albert et al., 2022] as well as A. Schrab, I. Kim, M. Albert, B. Guedj and A. Gretton [Schrab et al., 2023]. We are interested, on the one hand, in testing the independence of two vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ from the observation of an n-sample $((X_1, Y_1), \ldots, (X_n, Y_n))$ and, on the other hand, in testing that two independent samples of random variables with values in $\mathbb{R}^p$, $(X_1, \ldots, X_m)$ i. i.d. with probability distribution P and $(Y_1, \ldots, Y_n)$ i.i.d. with probability distribution Q, have the same distribution. These two papers have in common their use of the notion of MMD (Maximum Mean Discrepancy), which defines a metric between probability distributions based on kernels in reproducing kernel Hilbert spaces (RKHS). More precisely, given an RKHS $\mathcal{H}_k$ associated with kernel k, the MMD between two probability measures P and Q is defined by

$$\mathrm{MMD}(P, Q, \mathcal{H}_k) = \sup_{f \in \mathcal{H}_k, \|f\|_{\mathcal{H}_k} \leq 1} \mathbb{E}_{X \sim P}\left[f(X)\right] - \mathbb{E}_{Y \sim Q}\left[f(Y)\right].$$

For certain types of kernels, known as characteristic kernels, the nullity of $\mathrm{MMD}(P, Q, \mathcal{H}_k)$ is equivalent to the equality of the probability measures P and Q.

For the problem of testing the equality of the laws P and Q of two samples, we focus on estimating the quantity $\mathrm{MMD}(P, Q, \mathcal{H}_k)$, for a certain choice of kernel k, following the seminal work of [Gretton et al., 2007]. Furthermore, to test the independence of two random vectors X and Y, we propose to use the Hilbert-Schmidt independence criterion (HSIC) introduced by [Gretton et al., 2005]), which is none other than the MMD (associated with a certain kernel) between the law of the pair (X, Y) and the product of the marginal laws.

The aim of this presentation will be to propose estimators of the MMD and HSIC, and then to derive tests of homogeneity and independence. In particular, we will explain how the use of permutation techniques allows to guarantee the level of the tests. In addition, we will give power results for the tests, which are based on exponential inequalities for U-statistics due to [Arcones and Giné, 1993] and [Giné et al., 2000]. We will also discuss the choice of kernels used and show the benefits of aggregating tests associated with different kernels.

Keywords. Nonparametric tests, Maximum Mean Discrepancy, Hilbert-Schmidt Independence Criterion, permutation methods, aggregated tests.

Références

[Albert et al., 2022] Albert, M., Laurent, B., Marrel, A., and Meynaoui, A. (2022). Adaptive test of independence based on HSIC measures. *Ann. Statist.*, 50(2) :858–879.

[Arcones and Giné, 1993] Arcones, M. A. and Giné, E. (1993). Limit theorems for u-processes. *Ann. Probab.*, 21(3) :1494–1542.

[Giné et al., 2000] Giné, E., Latał a, R., and Zinn, J. (2000). Exponential and moment inequalities for U-statistics. In *High dimensional probability, II (Seattle, WA, 1999)*, volume 47 of *Progr. Probab.*, pages 13–38. Birkhäuser Boston, Boston, MA.

[Gretton et al., 2007] Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2007). A kernel method for the two-sample problem. In *Advances in Neural Information Processing Systems*, pages 513–520, Cambridge, MA. MIT Press.

[Gretton et al., 2005] Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. In Jain, S., Simon, H. U., and Tomita, E., editors, *Algorithmic Learning Theory*, pages 63–77, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Schrab et al., 2023] Schrab, A., Kim, I., Albert, M., Laurent, B., Guedj, B., and Gretton, A. (2023). MMD aggregated two-sample test. *J. Mach. Learn. Res.*, 24 :Paper No. [194], 81.