Nouveau test de normalité en grande dimension

Jérémie Kellner [1] & Alain Celisse [2]

[1] *Laboratoire de Mathématiques*
UMR 8524 CNRS - Université Lille 1 - MODAL team-project Inria
59655, Villeneuve d'Ascq Cedex
jeremie.kellner@inria.fr
[2] *Laboratoire de Mathématiques*
UMR 8524 CNRS - Université Lille 1 - MODAL team-project Inria
59655, Villeneuve d'Ascq Cedex

Résumé. Nous proposons un nouveau test d'adéquation à la loi normale dans un espace de Hilbert à noyau reproduisant (RKHS), qui peut se réduire au cas de la grande dimension par un choix de noyau approprié. Il partage des idées communes avec la *Maximum Mean Discrepancy* (MMD) tout en ayant un temps d'éxecution plus rapide et en étant applicable à un type de données plus large (graphes, séquences d'ADN,...). Nous proposons d'établir des résultats théoriques pour les erreurs de type I et II. Des simulations sur données artificielles et réelles illustrent l'amélioration apportée par notre test en comparaison d'autres approches en grande dimension.

Mots-clés. rkhs, noyaux, test d'adéquation, test de normalité, transformée de Laplace

Abstract. We propose a new goodness-of-fit test for normality in a Reproducing Kernel Hilbert Space (RKHS) which reduces to the high-dimensional case for an adequate kernel. It shares common ideas with the Maximum Mean Discrepancy (MMD) it outperforms both in terms of computation time and applicability to a wider range of data (graphs, DNA sequences,...). Theoretical results are derived for the Type-I and Type-II errors. Synthetic and real data also illustrate the practical improvement allowed by our test compared with other leading approaches in high-dimensional settings.

Keywords. rkhs, kernel, goodness-of-fit test, normality test, Laplace transform

1 Contexte et motivations

Traiter des données non-vectorielles telles que des séquences d'ADN requiert généralement l'usage d'un noyau [Aro50]. Toute procédure est ensuite menée dans l'espace de Hilbert à noyau reproduisant (RKHS) associé dans lequel les observations sont généralement supposées gaussiennes. Par example, [BFG12] propose une méthode de classification supervisée et non-supervisée en modélisant chaque classe par un processus gaussien. Cette

hypothèse-clé de normalité est souvent faîte implicitement, comme par example pour l'Analyse à Composantes Principales à noyau [Zwa05] afin de contrôler l'erreur de reconstruction [Nik10], ou encore dans [SKK13] où un test d'égalité de moyennes est employé dans un cas de grande dimension. Il semble alors essentiel de vérifier cette hypothèse cruciale.

Selon la structure du RKHS (dimension finie ou infinie), on peut faire appel à des tests de normalité de type Cramer-von-Mises [Mar70, HZ90, SR05]. Néanmoins, ces tests se révèlent moins puissants lorsque la dimension augmente [voir SR05, Table 3]. Une méthode alternative consiste à projeter des objets de grande dimension sur des espaces unidimensionnels choisis aléatoirement, puis à appliquer un test univarié sur ces marginales [CAFR06]. Toutefois, de telles approches pâtissent d'une perte de puissance [voir CAFR06, Section 4.2]. Plus spécifiquement dans le cas d'un RKHS, [GRSS07] a introduit la *Maximum Mean Discrepancy* (MMD) et proposé un test statistique pour distinguer la distribution de deux échantillons. Cette approche souffre cependant de plusieurs limitations : (*i*) le noyau doit être *caractéristique* [GBR$^+$12], (*ii*) le temps de calcul requis est important, (*iii*) elle nécessite plusieurs approximations qui réduisent la puissance du test.

Notre contribution principale est de fournir un test statistique d'adéquation à la loi normale qui soit algorithmiquement efficace et dédié à des types de données très généraux (représentées dans un RKHS potentiellement de dimension infinie).

2 Le test Laplace-MMD

2.1 Stratégie

Nous disposons d'un échantillon $(X_1, \ldots, X_n)$ à valeurs dans un ensemble $\mathcal{X}$ muni d'un noyau $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. À partir de k, on peut bâtir (sous certaines conditions) un espace de fonctions $H(k)$ composé de toutes les combinaisons linéaires des fonctions d'évaluation $k(x, .)$, $x \in \mathcal{X}$. $H(k)$ est le RKHS associé à k.

Considérons les variables indépendantes $Y_1, \ldots, Y_n$ où $Y_i = k(X_i, .)$ de distribution P dans $H(k)$. Notre but est de tester l'hypothèse nulle H_0 selon laquelle $(Y_1, \ldots, Y_n)$ est issu d'un processus gaussien $P_0 = \mathcal{N}(\mu, \Sigma)$. Une variable Z est gaussienne si toute marginale $< Z, f >_{H(k)}, f \in H(k)$ suit une loi gaussienne univariée. Pour ce faire, nous proposons de mesurer l'écart entre la loi de l'échantillon et la distribution nulle via une quantité que nous baptisons *Laplace-MMD*, définie par

$$\Delta L(P, P_0) = \sup_{f \in H(k), ||f|| \leq 1} \left| \mathbb{E}_{Y \sim P} e^{<Y,f>} - \mathbb{E}_{Z \sim P_0} e^{<Z,f>} \right|. \tag{1}$$

Elle s'appuie sur la différence entre les transformées de Laplace de toutes les marginales de Y et de Z, et ne garde que la plus grande de ces différences.

Le côté pratique de notre méthode est que (1) peut se réécrire comme une statistique entièrement calculable. En effet, en introduisant un second RKHS $H(\bar{k})$ associé au noyau

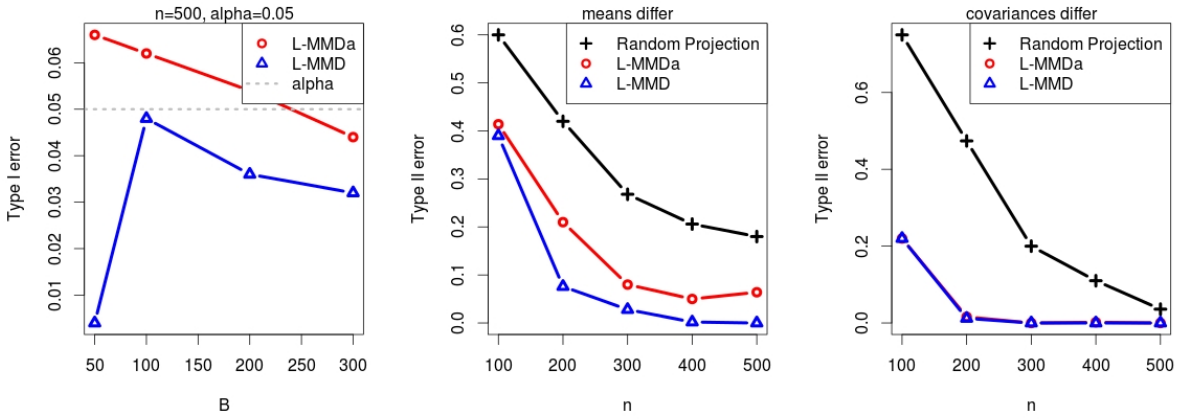

FIGURE 1 – **Gauche :** Erreurs de type-I pour les tests L-MMDa (• rouge) et L-MMD (Δ bleu). **Centre-Droite :** Erreurs de type-II pour les tests de Projection Aléatoire (+ noir), L-MMDa (• rouge) et L-MMD (Δ bleu). Centre : La distribution nulle et l'alternative diffèrent par leurs moyennes. Droite : La distribution nulle et l'alternative diffèrent par leurs covariances.

$\bar{k} = \exp(< ., . >_{H(k)})$ défini sur $H(k)$, on peut associer chaque distribution P sur $H(k)$ de façon unique à un élément $\bar{\mu}_P$ de $H(\bar{k})$. L'écart entre P et P_0 peut alors s'évaluer comme la différence entre les vecteurs correspondants $\bar{\mu}_P$ et $\bar{\mu}_{P_0}$.

Pour mettre en place le test, on estime la distribution de la statistique L-MMD sous H_0 via une méthode de Monte-Carlo. Cela permet de gagner en puissance par rapport à la startégie asymptotique exploitée par la MMD. De plus, notre méthode affiche également un temps de calcul de l'ordre de $\mathcal{O}(Bn^2)$ (B étant le nombre d'itérations de Monte-Carlo), plus rapide que $\mathcal{O}(n^3)$ pour la MMD tant que B est petit par rapport à n.

2.2 Performances

Les performances de L-MMD en termes d'erreur de type-I/II sont établies de façon théorique et empirique. Dans la Figure 1, elles sont comparées avec celles d'autres tests en grande dimension : Projection Aléatoire et un test que l'on appellera L-MMDa, qui utilise la stratégie asymptotique de la MMD dans la L-MMD.

Les erreurs de Type-I de L-MMD et L-MMDa sont représentées en fonction de B (qui représente le nombre de simulations Monte-Carlo pour estimer le quantile des distributions non-asymptotique et asymptotique). Le niveau de confiance est fixé à $\alpha = 0.05$. On observe alors que l'approche non-asymptotique respecte systématiquement le niveau de contrôle α, tandis que l'approche asymptotique le dépasse lorsque B n'est pas assez grand.

Pour évaluer l'erreur de Type-II, on considère la distribution nulle $P_0 = \mathcal{N}(\mu, \Sigma)$ et

deux alternatives $P_1 = \mathcal{N}(\mu_1, \Sigma)$ $(\mu_1 \neq \mu)$ et $P_2 = \mathcal{N}(\mu, \Sigma_2)$ $(\Sigma_2 \neq \Sigma)$. Dans les deux cas, L-MMD affiche une puissance supérieure à celles des tests L-MMDa et Projection Aléatoire.

Références

[Aro50] N. Aronszajn. Theory of reproducing kernels. May 1950.

[BFG12] C. Bouveyron, M. Fauvel, and S. Girard. Kernel discriminant analysis and clustering with parsimonious gaussian process models. 2012.

[CAFR06] J.A. Cuesta-Albertos, R. Fraiman, and T. Ransford. Random projections and goodness-of-fit test in infinite-dimensional spaces. *Bulletin of the Brazilian Mathematical Society*, pages 1–25, June 2006.

[GBR+12] A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, March 2012.

[GRSS07] K. Gretton, A.and Borgwardt, M. Rasch, B. Schoelkopf, and A. Smola. A kernel method for the two-sample-problem. In B. Schoelkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19 of *MIT Press, Cambridge*, pages 513–520, 2007.

[HZ90] N. Henze and B. Zirkler. A class of invariant and consistent tests for multivariate normality. *Comm. Statist. Theory Methods*, 19 :3595–3617, 1990.

[Mar70] K.V. Mardia. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57 :519–530, 1970.

[Nik10] S. Nikolov. Principal component analysis : Review and extensions. 2010.

[SKK13] M.S. Srivastava, S. Katayama, and Y. Kano. A two-sample test in high dimensional data. *Journal of Multivariate Analysis*, pages 349–358, 2013.

[SR05] G.J. Szekely and R.L. Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93 :58–80, 2005.

[Zwa05] L. Zwald. *Performances d'Algorithmes Statistiques d'Apprentissage : "Kernel Projection Machine" et Analyse en Composantes Principales à Noyaux.* 2005.