Test de normalité en grande dimension par méthodes à noyaux

Jérémie Kellner [1] & Alain Celisse [1]

[1] *Laboratoire de Mathématiques*
UMR 8524 CNRS - Université Lille 1 - MODAL team-project Inria
59655, Villeneuve d'Ascq Cedex

Résumé. Nous proposons un nouveau test de normalité dans un espace de Hilbert à noyau reproduisant (RKHS). Ce test reprend le principe de la MMD (Maximum Mean Discrepancy) - traditionnellement employé pour des tests d'homogénéité ou d'indépendance. Notre méthode intègre une procédure spéciale de bootstrap paramétrique - typique des tests d'adéquation - qui est parcimonieuse en temps de calcul par rapport au bootstrap paramétrique standard. En outre, une borne théorique pour l'erreur de Type-II est donnée. Enfin, des simulations montrent la puissance de notre test là où les tests de normalité courants deviennent rapidement inutilisables en grande dimension.

Mots-clés. RKHS, méthodes à noyaux, processus gaussien, test d'adéquation, test de normalité, bootstrap paramétrique, espace fonctionnel, grande dimension

Abstract. We propose a new goodness-of-fit test for normality in a Reproducing Kernel Hilbert Space (RKHS). This test thrives on the same principle as the MMD (Maximum Mean Discrepancy) which is usually used for homogeneity or independence testing. Our method makes use of a special kind of parametric bootstrap (typical of goodness-of-fit tests) which is more efficient than standard parametric bootstrap. Moreover, an upper bound for the Type-II error is provided. Experiments illustrate the practical improvement allowed by our test in high-dimensional settings where common normality tests are known to fail.

Keywords. RKHS, kernel methods, Gaussian process, goodness-of-fit test, normality test, parametric bootstrap, functionnal space, high-dimension

1 Processus gaussiens dans un RKHS

Les méthodes à noyaux permettent de traiter des types de données de natures diverses (séquences d'ADN, graphes, ...) en transposant le problème dans un contexte maniable. Formellement, étant données des observations dans un ensemble quelconque $\mathcal{X}$ et une fonction $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (le noyau), on peut associer à une observation X le vecteur $k(X, .)$ vivant dans un espace de fonctions engendré par k appelé le RKHS lié à k et noté $H(k)$ [Aro50]. On est donc ramené à manipuler des données vectorielles (en grande dimension).

Il est très courant que les observations dans le RKHS soient supposées gaussiennes. Par exemple, [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 faite implicitement, comme par exemple pour l'analyse en 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 nécessaire 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 directions unidimensionnelles choisies aléatoirement, puis d'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, [GBR+07] a introduit la *Maximum Mean Discrepancy* (MMD) et a proposé un test statistique pour distinguer la distribution de deux échantillons. La MMD a egalement été utilisée pour établir un test d'indépendance [GFT+07]. Néanmoins, le cas du test d'adéquation (de normalité par exemple) n'a pas encore été étudié dans ce contexte.

Notre contribution principale est de fournir un test statistique d'adéquation à la loi normale qui soit algorithmiquement efficace et pouvant être appliqué à des données dans un RKHS ($\mathbb{R}^d$ étant un cas particulier de RKHS). Notre test intègre une procédure de bootstrap paramétrique qui est algorithmiquement allégée par rapport à sa version standard.

2 Test de normalité dans un RKHS

2.1 Statistique de test

Nous disposons d'un échantillon $(X_1, \ldots, X_n)$ à valeurs dans un ensemble $\mathcal{X}$. $\mathcal{X}$ est supposé muni d'un noyau $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. On peut bâtir (sous certaines conditions) à partir de k un espace de fonctions $H(k)$ contenant toutes les combinaisons linéaires des fonctions d'évaluation $k(x, .)$, $x \in \mathcal{X}$. $H(k)$ est le RKHS associé à k [Aro50].

Considérons les variables $Y_i = k(X_i, .)$ $(i = 1, \ldots, n)$ 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 $P_0 = \mathcal{N}(\mu, \Sigma)$ (c'est-à-dire $< Y, f >_{H(k)} \sim \mathcal{N}(< \mu, f >_{H(k)}, < \Sigma f, f >_{H(k)})$ pour tout $f \in H(k)$) dont les paramètres μ (moyenne) et Σ (covariance) sont (en général) inconnus et doivent être estimés.

Pour ce faire, nous considérons un second noyau $\bar{k}$ défini sur $H(\bar{k})$ et nous associons à chaque distribution P sur $H(k)$ l'élément moyen $\bar{\mu}[P]$ dans $H(\bar{k})$ [BTA04, Chapitre 4]

$$P \mapsto \bar{\mu}[P] := \mathbb{E}_{Y \sim P} \bar{k}(Y, .) \ . \tag{2.1}$$

L'idée est de comparer deux distributions dans $H(k)$ au travers de leurs éléments moyens respectifs dans $H(\bar{k})$. Pour cela, il faut s'assurer que $\bar{k}$ soit tel que l'application (2.1) est injective (autrement dit, $||\bar{\mu}[P] - \bar{\mu}[Q]||_{H(\bar{k})} = 0$ implique $P = Q$). Un tel noyau $\bar{k}$ est dit caractéristique [SFL11]. Par exemple, il est bien connu que le noyau exponentiel $\bar{k} = \exp(< ., . >_{H(k)})$ et le noyau gaussien $\bar{k} = \exp(-\sigma||. - .||^2_{H(k)})$ $(\sigma > 0)$ sont caractéristiques [FSGS09].

Ainsi, dans notre cadre de test, la distance de P à la famille des lois normales est définie via la quantité

$$\Delta = \|\bar{\mu}[P] - \bar{\mu}[\mathcal{N}(\mu, \Sigma)]\|^2_{H(\bar{k})} \quad ,$$

où $\mu = \mathbb{E}_P Y$ et $\Sigma = \mathbb{E}_P (Y - \mu)^{\otimes 2}$.

N'ayant pas accès entièrement à la véritable distribution P, nous considérons en pratique la statistique de test

$$\hat{\Delta} = \left\| \bar{\mu}[\hat{P}] - \bar{\mu}[\mathcal{N}(\hat{\mu}, \hat{\Sigma})] \right\|^2_{H(\bar{k})} \quad , \tag{2.2}$$

où $\bar{\mu}[\hat{P}] = (1/n) \sum_{i=1}^n \bar{k}(Y_i, .)$ et $\hat{\mu}$ et $\hat{\Sigma}$ sont des estimateurs consistants de μ et Σ.

2.2 Une version rapide de bootstrap paramétrique

Pour mettre en place un test de normalité à partir de la statistique (2.2), il faut estimer sa distribution sous l'hypothèse nulle et en déduire une valeur critique liée à un niveau de confiance α. Il est possible de l'estimer avec une procédure de bootstrap paramétrique, qui consiste à générer B échantillons $(Y_1^{(b)}, \ldots, Y_n^{(b)})$ $(b = 1, \ldots, n)$ de la loi $\mathcal{N}(\hat{\mu}, \hat{\Sigma})$ puis de calculer B quantités $||\bar{\mu}[\hat{P}^{(b)}] - \bar{\mu}[\mathcal{N}(\hat{\mu}^{(b)}, \hat{\Sigma}^{(b)})]||^2_{H(\bar{k})}$ où $\hat{P}^{(b)}$, $\hat{\mu}^{(b)}$ et $\hat{\Sigma}^{(b)}$ sont recalculés à partir des $Y_1^{(b)}, \ldots, Y_n^{(b)}$. La distribution empirique donnée par ces B quantités permet d'approcher de façon consistante en n et B la véritable distribution nulle de $\hat{\Delta}$. Mais le fait de devoir recalculer la covariance empirique $\hat{\Sigma}^{(b)}$ à chaque réplication bootstrap alourdit les calculs (décomposition en valeur propres d'une matrice $n \times n$, soit une complexité de l'ordre de $\mathcal{O}(Bn^3)$).

Pour pallier cette limitation, nous avons adapté à notre cas une version de bootstrap paramétrique allégée en temps de calcul proposée dans [KY12]. L'idée générale consiste à scinder la différence de vecteurs à l'intérieur de la norme dans (2.2) en une partie comprenant la distance à la famille de loi normale et une seconde partie rendant compte de l'erreur d'estimation de μ et Σ. Plus précisément, étant donnés des multiplicateurs bootstrap (recentrés) $\bar{Z}_1^{(b)}, \ldots, \bar{Z}_n^{(b)}$ $(b = 1, \ldots, B)$, on considère

$$\hat{\Delta}^{(b)} = \left\| \bar{\mu}[\hat{P}^{(b)}] - D_{(\hat{\mu}, \hat{\Sigma})} \varphi(\hat{\mu}^{(b)}, \hat{\Sigma}^{(b)}) \right\|^2_{H(\bar{k})}, \quad b = 1 \ldots B \quad , \tag{2.3}$$

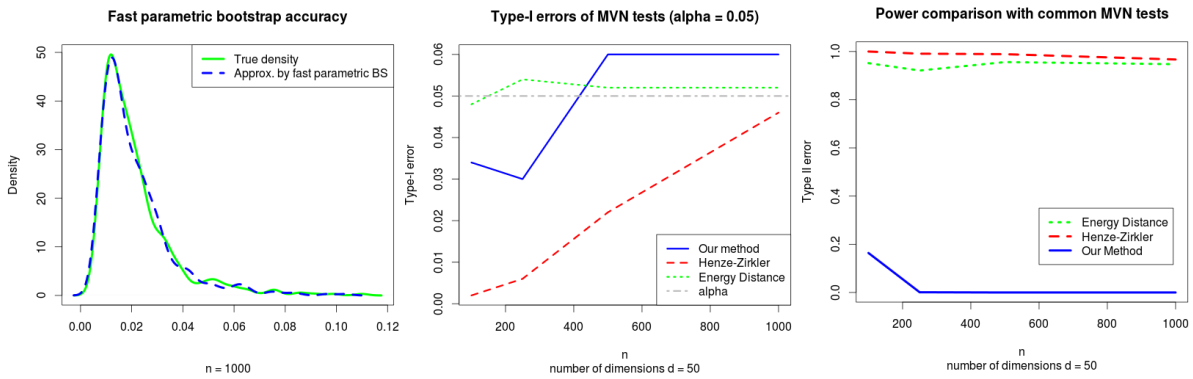

FIGURE 1 – (**Gauche :** Densités de la distribution de $\hat{\Delta}$ et de la distribution des replications bootstrap $\hat{\Delta}^{(b)}$, avec $n = 1000$; **Centre et droite :** comparaison des erreurs de type-I et type-II de notre test avec d'autres tests de normalité multivariée (MVN) : Henze-Zirkler et Energy Distance, avec le nombre de dimensions égal à $d = 50$.

où $\hat{P}^{(b)} = (1/n)\sum_{i=1}^{n} \bar{Z}_i^{(b)} \bar{k}(Y_i, .)$, $\hat{\mu}^{(b)} = (1/n)\sum_{i=1}^{n} \bar{Z}_i^{(b)} Y_i$, $\hat{\Sigma}^{(b)} = (1/n)\sum_{i=1}^{n} \bar{Z}_i^{(b)} (Y_i - \hat{\mu})^{\otimes 2}$ et $D\varphi$ représente la dérivée de Fréchet de $\varphi : (\mu, \Sigma) \mapsto \bar{\mu}[\mathcal{N}(\mu, \Sigma)]$. On obtient alors une estimation (consistante) de la valeur critique pour notre test avec une complexité algorithmique de l'ordre de $\mathcal{O}(Bn^2)$ au lieu de $\mathcal{O}(Bn^3)$.

La précision de cette méthode de bootstrap paramétrique se vérifie en pratique. Par exemple, la figure de gauche (Figure 1) illustre ceci dans le cas $n = 1000$ en superposant les densités de la distribution cible (celle de $\hat{\Delta}$) et de la distribution bootstrap (celle de $\hat{\Delta}^{(b)}$). Un test de Kolmogorov-Smirnov dans cette simulation renvoie une p-valeur de 0.978.

2.3 Performances

Les performances de notre tests sont établies thoériquement et empiriquement. Outre la validité asymptotique de notre procédure rapide de bootstrap, nous avons mis en évidence une borne exponentielle pour l'erreur de Type-II de la forme :

$$(1 + o_{n,B}(1)) \exp\left(-nL^2[C + o_n(1)]\right) \quad , \tag{2.4}$$

où $L^2 = ||\bar{\mu}[P] - \bar{\mu}[P_0]||^2_{H(\bar{k})}$ désigne l'écart à la distribution nulle et $C > 0$.

Les figures du centre et de droite (Figure 1) comparent l'efficacité de notre test en terme d'erreurs de type-I et de type II par rapport à plusieurs tests multivariés de normalité fréquemment employés : Henze-Zirkler [HZ90] et Energy Distance [SR05]. On fixe le niveau de confiance $\alpha = 0.05$ et la distribution alternative comme étant un mélange de deux gaussiennes de moyennes différentes (proportions de classes égales).

Avec un nombre de dimensions égal à $d = 50$, les tests concurrents ont une erreur de type II stagnant à 1, tandis que celle de notre test décroît très vite vers 0. On observe

que ces puissances de tests sont obtenues avec un niveau de confiance effectif (erreur de type-I) sensiblement égal pour tous les tests (d'autant plus que n est grand). Ainsi, la puissance de notre test par rapport aux autres n'est pas artificiellement induite par un niveau de confiance réel trop conservateur (erreur de type-I trop petite).

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.

[BTA04] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer, 2004.

[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.

[FSGS09] K. Fukumizu, B. Sriperumbudur, A. Gretton, and B. Schölkopf. Characteristic kernels on groups and semigroups. 2009.

[GBR+07] A. Gretton, K. Borgwardt, M. Rätsch, B. Schölkopf, 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.

[GFT+07] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B Schölkopf, and A. J. Smola. A kernel statistical test of independence. *NIPS*, 21, 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.

[KY12] I. Kojadinovic and J. Yan. Goodness-of-fit testing based on a weighted bootstrap : A fast large-sample alternative to the parametric bootstrap. *The Canadian Journal of Statistics*, 40 :3 :480–501, 2012.

[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.

[SFL11] B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12 :2389–2410, 2011.

[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.