Projections aléatoires dans des espaces à noyaux. Application à la detection d'outliers.

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 présentons un résultat nouveau traitant des projections d'une variable aléatoire plongée dans un espace à noyaux (ou RKHS). Plus précisément, dans le cas des noyaux de type RBF (pour Radial Basis Functions), nous montrons que la plupart de ces projections sont proches en loi d'une Gaussienne, pour peu que le noyau utilisé soit correctement renormalisé et son hyperparamètre suffisamment grand. Ce résultat permet de transformer des données de base via un noyau pour obtenir de nouvelles observations satisfaisant des conditions de normalité, et permettant ainsi d'appliquer des méthodes statistiques reposant sur ce genre d'hypothèse. En particulier, nous exposerons une méthode de détection d'observations atypiques (ou outliers) qui exploite cette transformation.

Mots-clés. Noyaux, RKHS, détection de points atypiques, modèles gaussiens, projections aléatoires

Abstract. We present a new result about finite dimensional projections of an embedded variable in an RKHS (Reproducing Kernel Hilbert Space). Namely in the case of RBF (Radial Basis Function) kernels, we show that most of these projections behave closely to a Gaussian, assuming that the kernel is adequately renormalized and that its hyperparameter is large enough. Thus this result allows to transform an initial dataset into a new dataset that satisfies normality conditions, hence the possibility of performing statistical methods relying on this kind of assumptions. As an illustration, we present a new method of novelty detection which is based on this transformation.

Keywords. Kernels, RKHS, novelty detection, Gaussian models, random projection

1 Projections aléatoires dans un RKHS

1.1 Contexte et résultats pré-existants

Considérons une variable aléatoire X à valeurs dans un ensemble quelconque $\mathcal{X}$. Il peut s'agir d'un vecteur aléatoire avec $\mathcal{X} = \mathbb{R}^p$, ou encore d'un graphe aléatoire, d'une séquence d'ADN, d'un stream audio, etc. Il existe un certain nombre d'approches qui consistent

à considérer une transformation $\phi(X)$ pour pouvoir appliquer un modèle gaussien; citons comme exmples d'application le clustering [1] ou encore la détection d'anomalies [4]. On se pose la question suivante: est-il possible de construire une application ϕ telle de $\phi(X)$ suive - du moins approximativement - une distribution de type gaussienne?

Une telle transformation peut être définie via un noyau $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ semi-défini positif. Une telle fonction est souvent définie en tant que mesure de comparaison entre éléments de $\mathcal{X}$; par exemple, si x et y sont deux chaînes de caractères, on peut définir $k(x, y)$ comme étant le nombre de sous-chaînes communes partagées par x et y. Pour un tel k, il existe une application caractéristique (ou *feature map*) ϕ et un espace à noyau (ou *feature space*) $\mathcal{H}$ muni d'un produit scalaire $\langle ., . \rangle_{\mathcal{H}}$ tel que pour tout $x, y \in \mathcal{X}$, la propriété de reproduisance $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$ soit vérifiée.

L'intérêt de passer par une transformation ϕ est de forcer nos données à vérifier une certaine propriété requise. Par exemple, dans le cas des *Support Vector Machines* (SVM) à noyaux, le fait de plonger des observations séparées en deux classes dans un RKHS - qui est typiquement de grande dimension - permet à l'arrivée d'obtenir deux classes de points linéairement séparables. De même, nous exploitons la grande dimensionnalité du RKHS pour faire apparaître des gaussiennes dans l'espace à noyau. En effet, si l'on s'intéresse au cas multivarié $\mathbb{R}^D$ où D est grand, il existe des résultats [3, 2] selon lesquels la plupart des projections d'un vecteur aléatoire D-varié suivent approximativement une loi typique, qui s'avère être une loi gaussienne dans certains cas. Par exemple, si on note $X_1, \ldots, X_n$ un échantillon de n vecteurs de $\mathbb{R}^D$ et $\nu^T X_1, \ldots, \nu^T X_n$ une projection de cet échantillon avec $\nu \in \mathcal{S}^{D-1}$ (où $\mathcal{S}^{D-1}$ désigne la sphère unité de $\mathbb{R}^D$), alors [3] montre que si $X_1, \ldots, X_n$ est approximativement orthonormal, alors l'échantillon $\nu^T X_1, \ldots, \nu^T X_n$ est proche d'une loi normale quand $n, D \to +\infty$, et ce pour la plupart des directions ν. [2] généralise cette propriété pour des projections de dimension $p \geqslant 1$ de la manière suivante: si X est un vecteur aléatoire de $\mathbb{R}^D$ et $\Theta \in \mathcal{M}_{p,D}(\mathbb{R})$ une matrice de projection, alors la distribution de $\Theta^T X$ tend à être proche d'un *scale-mixture* de gaussiennes σG où $G \sim \mathcal{N}(0, I_p)$ et σ est une copie indépendante de $\|X\|/\sqrt{D}$. La distance en variation totale entre les deux distributions est de l'ordre de

$$\tilde{O}\left(\frac{\mathrm{ecc}(X)p}{D}\right)^{1/4} ,$$

où $\mathrm{ecc}(X)$ mesure l'excentricité de X, c'est-à-dire le degré d'aplatissement du spectre de covariance de X (avec une valeur minimale si X est un vecteur isotropique). Ainsi, la convergence est assurée lorsque D est grand et $\mathrm{ecc}(X)$ est relativement petite.

1.2 Nouveau résultat

Dans la suite, nous considérons la famille des noyaux dits *Radial Basis Function* (RBF) $k_\gamma(., .)$ paramétrée par γ et définie par

$$k_\gamma(x, y) = \exp\left(-\gamma \|x - y\|^2\right) .$$

Notons $\mathcal{H}_\gamma$ le RKHS associé à k_γ et $\langle .,. \rangle_\gamma$ son produit scalaire.

Deux points sont à remarquer concernant cette catégorie de noyaux:

1 Les RKHS $\mathcal{H}_\gamma$ correspondants aux $k_\gamma(.,.)$ sont emboîtés, autrement dit on a $\mathcal{H}_{\gamma_1} \subseteq \mathcal{H}_{\gamma_2}$ pour tout $\gamma_1 < \gamma_2$;

2 Si l'on note ϕ_γ le *feature map* associé à $k_\gamma(.,.)$, alors les vecteurs $\phi_\gamma(x)$ pour $x \in \mathcal{X}$ sont approximativement orthonormaux lorsque $\gamma \to +\infty$, c'est-à-dire que pour tout $x, y \in \mathcal{X}$,

$$\langle \phi_\gamma(x), \phi_\gamma(y) \rangle_\gamma = k_\gamma(x,y) \underset{\gamma \to +\infty}{\longrightarrow} \mathbb{1}_{x=y} \ .$$

Le premier point nous dit que le paramètre γ contrôle la taille du RKHS, de la même manière que D contrôle la taille de $\mathbb{R}^D$ dans le cas multivarié. Enfin, le deuxième point indique que l'excentricité des points dans le RKHS tend à être minimale lorsque $\gamma \to +\infty$. Ainsi, les résultats connus dans le cas $\mathbb{R}^D$ laissent espérer un comportement similaire pour le RKHS d'un noyau RBF.

Nous considérons dans la suite $P_h(X)$ la projection de $\phi_\gamma(X)$ sur le sous-espace de $\mathcal{H}_\gamma$ engendré par p vecteurs aléatoires $h_1, \ldots, h_p$:

$$P_h(X) = (\langle \phi_\gamma(X), h_1 \rangle_\gamma, \ldots, \langle \phi_\gamma(X), h_p \rangle_\gamma)^T \ .$$

On suppose dans la suite que les vecteurs $h_1, \ldots, h_p$ ont été générés indépendamment par un processus gaussien de moyenne nulle et ayant pour fonction de covariance $\Sigma_\gamma(x, x') = \mathbb{E} k_\gamma(X, x) k_\gamma(X, x')$, où $x, x' \in \mathcal{X}$.

Nous affirmons que, similairement au cas multivarié, $P_h(X)$ est proche d'un *scale-mixture* de gaussiennes σG lorsque γ tend vers l'infini. Dans notre cas, σ est une copie de $\sqrt{\Sigma_\gamma(X,X)} = \sqrt{\mathbb{E}_{X'} k_\gamma^2(X, X')}$ et $G \sim \mathcal{N}(0, I_p)$ est indépendant de σ. Plus encore, nous montrons que σ est égal presque surement (et pas seulement en loi) à $\sqrt{\Sigma_\gamma(X,X)}$. Autrement dit, on peut vérifier que la projection renormalisée

$$\tilde{P}_h(X) = \Sigma_\gamma^{-1/2}(x,x) P_h(X) \ ,$$

est proche d'une simple gaussienne lorsque γ est grand.

Theorème 1.1. *Supposons que $\mathcal{X} = \mathbb{R}^D$ où $\|.-.\|$ désigne la distance euclidienne usuelle. Soit f la densité de X qui est supposée continue et bornée. On suppose également que le support de X a une mesure de Lebesgue finie $\mu_\mathcal{R}$. Alors, avec probabilité supérieure à $1-\delta$ par rapport aux $h_1, \ldots, h_p$,*

$$\Delta \leqslant \frac{C p^{3/2} \|f\|_4 \|f\|_2^{-1}}{\delta \gamma^{D/8}} \qquad et \qquad \tilde{\Delta} \leqslant \frac{\tilde{C} p^{3/2} \|f\|_4 \mu_\mathcal{R}^{1/2}}{\delta \gamma^{D/8}} \ ,$$

où Δ (resp. $\tilde{\Delta}$) est une mesure de l'écart entre les distributions de $P_h(X)$ et σG (resp. $\tilde{P}_h(X)$ et $\mathcal{N}(0, I_p)$), et $C, \tilde{C}$ sont des constantes.

2 Application à la détection de points atypiques

2.1 Principe

Supposons que l'on dispose d'une séquence d'observations (à valeurs dans $\mathcal{X}$)

$$X_1, X_2, \ldots, X_n, x$$

où les n points précédents $X_1, \ldots, X_n$ sont supposés être des réalisations indépendantes d'une distribution P. On appelle ces n points des *inliers*. Le but est de tester le nouveau point x afin de déterminer s'il a été aussi généré par P; dans le cas contraire, on dit que x est un *outlier*. Notons $H_0 : x \sim P$ l'hypothèse nulle à tester, et $H_A : x \sim Q, P \neq Q$ l'hypothèse alternative.

Pour que le problème soit faisable, on doit supposer que la distribution P des *inliers* soit presque sûrement (ou avec grande probabilité) contenue dans une région $\mathcal{R}$ de $\mathcal{X}$, de sorte qu'un *outlier* soit défini comme une observation qui se situe en dehors de $\mathcal{R}$. Au final, le problème réside dans le fait d'estimer $\mathcal{R}$.

Notre approche de ce problème repose sur l'observation suivante: pour toute paire de points distincts $y, y^{'} \in \mathcal{X}$, on a $\langle \phi_\gamma(y), \phi_\gamma(y^{'}) \rangle_\gamma = k_\gamma(y, y^{'}) \to 0$ quand $\gamma \to +\infty$. Ainsi, si x n'appartient pas au support de P, ϕ_γ tend à être orthogonal au sous-espace $\text{Span}\{k_\gamma(y, .), y \in \mathcal{R}\}$ de $\mathcal{H}_\gamma$. Cela implique que la projection $P_h(x)$ est proche de 0 lorsque γ est grand, puisque $h_1, \ldots, h_p$ appartiennent à $\text{Span}\{k_\gamma(y, .), y \in \mathcal{R}\}$ presque surement.

De ce fait, la quantité $S(x) = \|P_h(x)\|^2_{\mathbb{R}^p}$ a deux comportements distincts lorsque $\gamma \to +\infty$:

- Si $x \in \mathcal{R}$, alors d'après le theorème 1.1, $S(x)$ converge en loi vers une variable $\sigma^2 V^2$ où σ^2 est une copie de $\Sigma_\gamma(X, X)$ et où $V^2 \sim \chi_2(p)$ est indépendant de σ^2;
- Si $X \notin \mathcal{R}$, alors $S(x) \to 0$ presque surement.

En pratique, le terme de covariance $\Sigma_\gamma(x, x)$ est remplacé par son estimateur empirique $\Sigma_{\gamma,n}(x, x) = (1/n) \sum_{i=1}^n k^2_\gamma(X_i, x)$. On note $P_{h,n}(x)$ la version empirique de $P_h(x)$ où $h_1, \ldots, h_p$ sont générés indépendamment par un processus gaussien de moyenne nulle et de covariance $\Sigma_{\gamma,n}$. La statistique de test devient alors

$$S_n(x) = \|P_{h,n}(x)\|^2_{\mathbb{R}^p} = n^{-2} \mathbf{k_x}^T H H^T \mathbf{k_x} \ ,$$

où $\mathbf{k_x} = (k_\gamma(X_1, x) \ldots k_\gamma(X_n, x))^T$ et où les entrées de $H \in \mathcal{M}_{n,p}(\mathbb{R})$ sont des $\mathcal{N}(0, 1)$ indépendantes. Étant donné α un niveau de contrôle voulu de l'erreur de Type-I, on définit un seuil τ_α qui vérifie $\mathbb{P}(\sigma^2 V^2 < \tau_\alpha) = \alpha$ et on rejettera H_0 si et seulement si $S_n(x) < \tau_\alpha$.

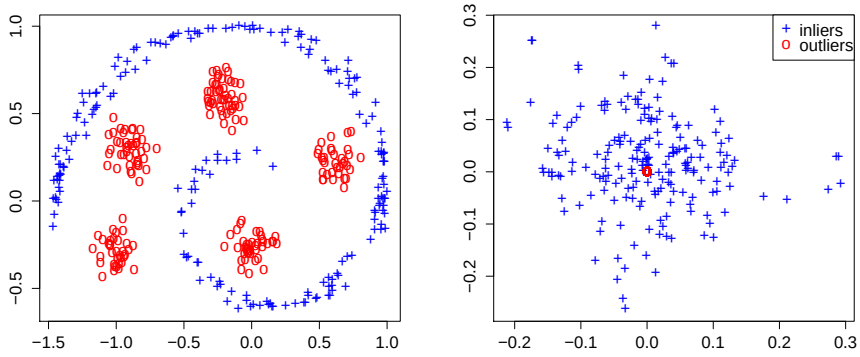

Figure 1: Jeu de données "spirale" (gauche), séparé en *inliers* (bleu, croix) et *outliers* (rouge, cercle). Le graphe de droite montre un exemple de projection de ces données dans l'espace à noyau pour $p = 2$ et $\gamma = 250$. À noter que les d'*outliers* (rouge) à droite sont concentrés autour de l'origine.

2.2 Simulations

On applique la méthode de détection de points atypiques présentée dans la section 2.1 au jeu de données représenté dans la figure 1. Les *inliers* sont distribués sur une région de $\mathbb{R}^2$ formant une spirale, et les *outliers* sont générés en dehors de cette spirale. Plus précisément, on génère un *training set* de $n = 500$ *inliers* (qui jouent le rôle des $X_1, \ldots, X_n$ selon la notation de la section 2.1), puis un *test set* de 200 *inliers* et 200 *outliers* (qui jouent le rôle du x dans la section 2.1). On teste les 400 observations du *test set* et on observe la proportion d'*inliers* rejetés à tort (erreur de Type-I) et la proportion d'*outliers* non-rejetés à tort (erreur de Type-II). On réitère le procédé pour différentes valeurs des paramètres γ et p, allant de 1 jusqu'à 1000 pour γ et de 2 jusqu'à 500 pour p. Les résultats obtenus sont donnés dans les tableaux de la figure 2.

Pour ce qui est de l'erreur de Type-I, on observe que dans une première phase (γ compris entre 1 et 100), celle-ci est proche de la valeur prescrite (ici $\alpha = 0.05$) , ce qui est conforme au théorème 1.1. En revanche, lorsque γ prend une valeur plus grande comme $\gamma = 1000$, l'erreur de Type-I n'est plus tout contrôlée. Cela est dû au fait qu'en pratique, le terme de covariance $\Sigma_\gamma(x, x)$ est remplacé par son estimateur empirique $\Sigma_{\gamma,n}(x, x)$. Or, on peut montrer via une inégalité de concentration de type Bennett que $\left|\Sigma_\gamma(x, x)\Sigma^{-1}_{\gamma,n}(x, x) - 1\right| = \mathcal{O}\left(\gamma^{D/4} n^{-1/2}\right)$. Ainsi il faut disposer d'un nombre suffisant n d'*inliers* antécédents pour pouvoir obtenir un bon contrôle de l'erreur de Type-I, et il faut aussi qu'à n fixé, γ ne soit pas trop grand par rapport à n. Quant à l'erreur de Type-II, on observe que celle-ci tend vers 0 lorsque $\gamma \to +\infty$. Ainsi, le calibrage de p et γ semble lié principalement au contrôle de l'erreur de Type-I et donc à la connaissance de la distribution des *inliers*, qui est accessible puisqu'on dispose en général d'un nombre important

	$\gamma = 1$	25	100	1000
$p = 2$	0.047	0.052	0.056	0.175
10	0.045	0.054	0.059	0.369
50	0.053	0.049	0.063	0.456
500	0.049	0.048	0.062	0.469

	$\gamma = 1$	25	100	1000
$p = 2$	0.962	0.078	0.006	0
10	0.951	0.006	0	0
50	0.910	0	0	0
500	0.841	0	0	0

Figure 2: Erreurs de Type-I (gauche) et de Type-II (droite) pour différentes valeurs de γ et p

d'*inliers* contrairement aux *outliers*. Une méthode pour calibrer de façon efficiente ces deux paramètres est l'objet de développements futurs.

References

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

[2] S. Dasgupta, D. Hsu, and N. Verma. A concentration theorem for projections. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2006.

[3] P. Diaconis and D. Freedman. Asymptotics of graphical projection pursuit. *The annals of statistics*, pages 793–815, 1984.

[4] V. Roth. Kernel Fisher Discriminants for Outlier Detection. *Neural Computation*, 18(4):942–960, 2006.