Un test statistique pour la détection d'anomalies basé sur l'erreur de reconstruction de l'ACP à noyau

Chloé Friguet[1] & Laetitia Chapel[1]

[1] *Univ. Bretagne-Sud, UMR 6074, IRISA, F-56000 Vannes, France*
chloe.friguet@univ-ubs.fr - laetitia.chapel@univ-ubs.fr

Résumé. La détection d'anomalies concerne l'identification de points dont le comportement dévie d'un modèle dit nominal. On présente ici un test statistique non paramétrique pour la détection d'anomalies dont la statistique de test est basée sur la distance entre un point à tester, représenté dans un espace de redescription, et sa projection sur l'espace généré par une Analyse en Composantes Principales à noyau réalisée sur un échantillon tiré d'une loi de probabilité nominale. La méthode est testée sur des données réelles et artificielles, et montre de bonnes performances en ce qui concerne à la fois de l'erreur de type-I et de type-II par rapport à des méthodes usuelles de détection d'anomalies. Cette communication est basée sur un article récemment publié par les mêmes auteurs [3].

Mots-clés. Détection d'anomalies, ACP à noyau, erreur de reconstruction

Abstract. Anomaly detection aims at declaring a query point as "normal" or not with respect to a nominal model. A non-parametric statistical test that allows the detection of anomalies given a set of (possibly high dimensional) sample points drawn from a nominal probability distribution is presented. Its test statistic is based on the distance between a query point, mapped in a feature space, and its projection on the eigen-structure of the kernel matrix computed on the sample points. The method is tested on both artificial and benchmarked real data sets and demonstrates good performances regarding both type-I and type-II errors *w.r.t.* competing methods. This communication is based on a recently published paper by the same authors [3].

Keywords. Anomaly detection, Kernel-PCA, reconstruction error

Support. Ce travail a été partiellement financé par le projet ANR Asterix (ANR-13-JS02-0005-01).

1 Introduction

La détection d'anomalies [1] a pour but de déclarer un point à tester $\boldsymbol{\eta}$ comme "normal" ou non par rapport à un modèle dit nominal. La distribution de probabilité nominale sous-jacente $f_0(\boldsymbol{x})$ sur l'espace $\mathcal{X} \subset \mathbb{R}^d$ est inconnue mais on dispose d'un ensemble de n points

nominaux *i.i.d.* $\mathcal{S} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$. Le problème de la détection d'anomalies peut être reformulé comme un test statistique :

$$\begin{cases} \mathcal{H}_0 : & \boldsymbol{\eta} \sim f_0 \quad \text{i.e. } \boldsymbol{\eta} \text{ est issu de la distribution nominale} \\ \mathcal{H}_1 : & \boldsymbol{\eta} \nsim f_0 \quad \text{i.e. } \boldsymbol{\eta} \text{ n'est pas issu de la distribution nominale} \end{cases}$$

qui permet de contrôler l'erreur de type-I à un niveau α fixé. Si la probabilité critique du test $p(\boldsymbol{\eta}) \leq \alpha$, l'hypothèse nulle est rejetée et $\boldsymbol{\eta}$ est déclaré comme étant une anomalie.

L'approche classique en détection d'anomalies consiste à déclarer comme telle un point $\boldsymbol{\eta}$ qui se trouve dans une zone où la densité est faible, en fixant un seuil t pour la densité nominale f_0 : $\boldsymbol{\eta}$ est considéré comme anomalie si $f_0(\boldsymbol{\eta}) < t$. Ainsi, la détection d'anomalies est liée à l'estimation de densité : si $\boldsymbol{\eta}$ est issu de la distribution nominale, on s'attend à ce qu'il se trouve au delà du seuil t avec une probabilité $p(\boldsymbol{\eta}) = 1 - F_0(t)$. En pratique, on évalue la probabilité critique en considérant le positionnement du point testé $\boldsymbol{\eta}$ par rapport à l'ensemble des points nominaux de l'échantillon $\mathcal{S}$:

$$\hat{p}(\boldsymbol{\eta}) = \frac{1}{n} \sum_{k=1}^{n} \Big(\mathbb{1}\{F(\boldsymbol{\eta}) \leq F(\boldsymbol{x}_k)\} \Big) \tag{1}$$

où $\mathbb{1}\{.\}$ est la fonction indicatrice. F représente un estimateur de la fonction de répartition nominale F_0.

Nous utilisons ici une approche non paramétrique, basée sur une mesure de substitution de la densité nominale, calculée à partir de l'erreur de reconstruction dans la projection du point testé dans un espace engendré par une ACP à noyau. Nous montrons que l'utilisation de cette mesure, initialement développée par [4], ainsi qu'une nouvelle définition de la probabilité critique, permet de controller l'erreur de type-I, tout en minimisant l'erreur de type-II.

2 Test de détection d'anomalie basé sur l'erreur de reconstruction de l'ACP à noyau

ACP à noyau et erreur de reconstruction Les méthodes de décomposition en valeurs singulières comme l'ACP ont pour objectif d'identifier et d'extraire la structure linéaire des données. Pour le cas des structures non-linéaires, des versions à noyau de ce type de méthodes ont été développées. Il s'agit d'appliquer l'ACP sur les données auxquelles on a appliqué une transformation non linéaire $\psi : \boldsymbol{x} \in \mathcal{X} \mapsto \psi(\boldsymbol{x}) \in \mathcal{F}$. On peut calculer les produits scalaires dans l'espace transformé sans jamais avoir à faire la transformation explicitement : c'est l'astuce du noyau $\kappa(\boldsymbol{x_i}; \boldsymbol{x_j}) = \langle \psi(\boldsymbol{x_i}), \psi(\boldsymbol{x_j}) \rangle$. La matrice de Gram $\boldsymbol{K}$ $\boldsymbol{K}_{ij} = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $\forall \boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{X}$ peut alors se décomposer en $\boldsymbol{K} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T$, où $\boldsymbol{\Lambda}$ représente la matrice diagonale des valeurs propres et $\boldsymbol{V} = \Big[\boldsymbol{\phi}_i\Big]_{1 \leq i \leq \infty}$ est la matrice des vecteurs propres.

Pour un point $\boldsymbol{\eta}$, on peut calculer sa projection dans l'espace engendré par l'ACP à noyau à partir de $\psi(\mathcal{S})$ [7] : $r(\boldsymbol{\eta} \rightarrow \psi(\mathcal{S})) = \kappa(\mathcal{S};\boldsymbol{\eta}) \cdot \boldsymbol{V} \cdot \Lambda^{-1/2}$ avec $\kappa(\mathcal{S};\boldsymbol{\eta}) = \Big[\kappa(\boldsymbol{x}_1;\boldsymbol{\eta});\ldots;\kappa(\boldsymbol{x}_n;\boldsymbol{\eta})\Big]$. La qualité de cette projection se mesure par l'erreur de reconstruction à travers la norme 2 des résidus $\tau_{\mathcal{S}}(\boldsymbol{\eta})$:

$$\tau_{\mathcal{S}}(\boldsymbol{\eta}) \quad = \quad \|\psi(\boldsymbol{\eta}) - r(\boldsymbol{\eta} \rightarrow \psi(\mathcal{S}))\|^2 . \tag{2}$$

On note ainsi que si $\boldsymbol{\eta} \subset \mathcal{S}$, $\tau_{\mathcal{S}}(\boldsymbol{\eta}) = 0$.

La qualité de la projection d'un point issu de la même distribution f_0 que $\mathcal{S}$ permet d'évaluer la performance de l'ACP à noyau et a été discuté dans la littérature [8, 7]. Ainsi, sous $\mathcal{H}_0$, $\tau_{\mathcal{S}}(\boldsymbol{\eta}) \rightarrow 0$ avec une grande probabilité, à condition que les valeurs propres décroissent rapidement. Dans ce cas, $\tau(\boldsymbol{\eta})$ est correctement estimé par $\tau_{\mathcal{S}}(\boldsymbol{\eta})$ et $\tau_{\mathcal{S}}(\boldsymbol{\eta}) \rightarrow 0$ lorsque n devient grand [7].

Lien entre estimation de densité et ACP à noyau La densité nominale f_0 est reliée à la fonction noyau $\kappa(\boldsymbol{x},\boldsymbol{y})$ par une approximation [8] à partir de l'échantillon $\mathcal{S}$:

$$\int_{\mathcal{X}} \kappa(\boldsymbol{x},\boldsymbol{y}) f_0(\boldsymbol{x}) \phi_i(\boldsymbol{x}) d\boldsymbol{x} \backsimeq \frac{1}{n} \sum_{k=1}^{n} \kappa(\boldsymbol{x}_k;\boldsymbol{y}) \phi_i(\boldsymbol{x}_k). \tag{3}$$

Ainsi, la décomposition de la matrice de Gram par ACP à noyau est reliée à la densité nominale, et nous proposons donc, plutôt que d'estimer directement f_0, d'utiliser l'erreur de reconstruction définie en (2) comme substitut pour F_0 dans (1). En effet, si le point testé $\boldsymbol{\eta}$ est issu de la densité nominale f_0, $\psi(\boldsymbol{\eta})$ et $\psi(\mathcal{X})$ sont très proches et donc l'erreur de reconstruction $\tau(\boldsymbol{\eta})$ est faible. Au contraire, si $\boldsymbol{\eta}$ n'est pas issu de la densité nominale, son erreur de reconstruction sera élevée. Ce comportement fait de $\tau(.)$ une mesure adequat pour positionner le point testé : seuiller τ est très similaire à un seuillage de la densité nominale.

Définition du test Pour définir la fonction de répartition empirique G de τ, on propose une méthode basée sur du ré-échantillonnage. À partir des observations $\mathcal{S}$

$$G(t) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}(\tau_{\mathcal{S}}^{-k}(\boldsymbol{x}_k) \leq t) \tag{4}$$

où $\tau_{\mathcal{S}}^{-k}(\boldsymbol{x}_k) = \kappa(\boldsymbol{x}_k;\boldsymbol{x}_k) - \kappa(\boldsymbol{x}_k;\mathcal{S}) \cdot \boldsymbol{K}(\mathcal{S}\backslash\{\boldsymbol{x}_k\})^{-1} \cdot \kappa(\boldsymbol{x}_k;\mathcal{S})^T$ est l'erreur de reconstruction de $\boldsymbol{x}_k$ calculée à partir de $\mathcal{S}\backslash\{\boldsymbol{x}_k\}$. La loi forte des grands nombres assure la convergence vers G_0 la fonction de répartition nominale pour tout t : $\sup_{t\in\mathbb{R}} |G(t) - G_0(t)| \mapsto 0$.

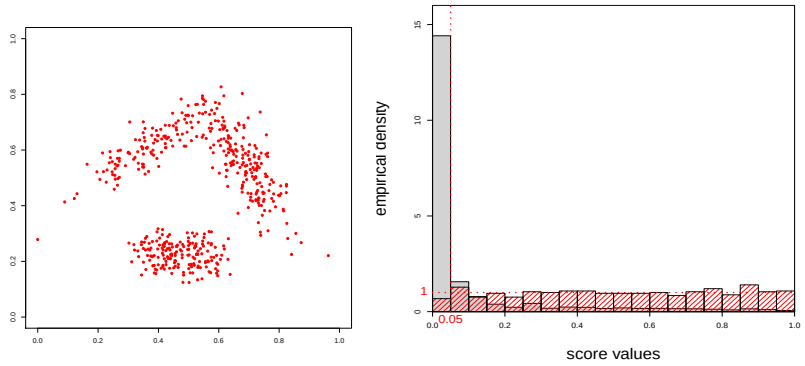
La probabilité critique du test est alors définie par :

$$\hat{p}(\boldsymbol{\eta}) \quad = \quad \frac{1}{n} \sum_{\boldsymbol{x}_k \in \mathcal{S}} \mathbb{1}\left\{\tau_{\mathcal{S}}(\boldsymbol{\eta}) \leq \tau_{\mathcal{S}}^{-k}(\boldsymbol{x}_k)\right\}. \tag{5}$$

Le calcul de tous les termes $\tau_{\mathcal{S}}^{-k}(\boldsymbol{x}_k)$ a une complexité en $O(n^3+n^2)$ et le test n'implique que des produits matriciels. En particulier, le calcul de l'inverse pour chaque matrice $\boldsymbol{K}(\mathcal{S}\setminus\{\boldsymbol{x}_k\})^{-1}$, $\boldsymbol{x}_k \in \mathcal{S}$ de taille $(n-1)\times(n-1)$ n'est pas nécessaire en considérant à chaque fois le complément de Schur à partir de $\boldsymbol{K}(\mathcal{S})^{-1}$. Afin d'éviter des problèmes numériques lors de ces calculs, ou afin de diminuer la sensibilité au bruit, des techniques de régularisation peuvent également être mises en œuvre.

3 Illustration de la méthode

Données artificielles Les performances de la méthode proposée sont évaluées sur un jeu de données artificiel, composé de 500 points nominaux (Fig. 1(a)). La Fig. 1(b) représente la distribution empirique des probabilités critiques obtenues pour 1 000 points nominaux d'une part, qui sont distribués uniformément, et 1 000 points tirés selon une loi uniforme bivariée, qui sont concentrés proche de 0.



(a) (b)

FIGURE 1 – (a) Echantillon nominal $\mathcal{S}$ utilisé pour le calcul de τ (b) Distribution des probabilités critiques (rouge : points nominaux - gris : points tirés selon une loi uniforme)

Données réelles On compare les performances de l'algorithme proposé (re-kpca) avec celles obtenues par un one-class SVM [6] et k-lpe [9], sur des données réelles ([5] et [2]). On choisit une des classes comme nominale et on y tire aléatoirement n observations. Les trois premiers jeux de données ont un nombre modéré de dimensions (moins de 10), tandis que les trois derniers ont un nombre de dimensions supérieur ou égal à 60. Pour évaluer les méthodes, on utilise l'aire sous la courbe ROC (AUC), et les erreurs de type-I et II (sauf pour one-class SVM pour qui il n'y a pas de méthode pour contrôler ces erreurs). Les résultats sont reportés dans les Tab. 1 et 2

One-class SVM présente les AUC les plus faibles. Les 2 autres méthodes ont des performances similaires en terme d'AUC et d'erreur de type-II pour les données de dimension modérée. Par contre, notre méthode permet d'obtenir de meilleurs résultats pour des dimensions élevées : on remarque que k-lpe, qui est basée sur les plus proches voisins, présente parfois des résultats incohérents pour les erreurs de type-I, et que ses performances sont dégradées en ce qui concerne les erreurs de type-II.

TABLE 1 – AUC (%) - Moyenne sur 100 repetitions

Dataset	$n = 50$			$n = 100$			$n = 500$		
	re-kpca	k-lpe	oc-svm	re-kpca	k-lpe	oc-svm	re-kpca	k-lpe	oc-svm
Banana	**88.04**	87.90	83.50	**90.47**	89.77	86.82	**92.53**	92.48	91.54
Diabetes	**73.21**	73.44	67.66	74.11	**74.75**	67.36	-	-	-
Thyroid	**98.15**	97.94	96.12	**99.04**	98.52	97.10	-	-	-
Mushroom	**97.80**	97.23	75.35	**99.04**	98.68	84.48	**99.83**	99.58	94.52
Sonar	**70.34**	64.86	60.77	**73.12**	70.34	61.79	-	-	-
USPS	**97.71**	96.07	90.50	**97.95**	97.05	93.49	**98.73**	97.80	95.96

TABLE 2 – Erreurs de type-I et II (%) - moyenne sur 100 répétitions

Dataset	α	**type-I**						**type-II**					
		$n = 50$		$n = 100$		$n = 500$		$n = 50$		$n = 100$		$n = 500$	
		re-kpca	k-lpe	re-kpca	k-lpe	re-kpca	k-lpe	re-kpca	k-lpe	re-kpca	k-lpe	re-kpca	k-lpe
Banana	2%	1.91	1.35	1.75	1.86	1.91	1.87	**74.53**	81.38	**64.51**	71.63	**45.18**	50.77
	5%	6.09	5.17	4.72	5.11	4.77	4.91	**44.21**	47.92	**38.82**	42.52	**28.49**	30.00
	20%	22.02	20.77	20.30	21.01	20.01	20.10	**17.48**	18.27	15.15	**14.57**	11.77	**11.47**
Diabetes	2%	1.51	1.52	2.04	1.92	-	-	95.18	**95.12**	**93.79**	94.49	-	-
	5%	5.28	5.77	5.02	4.70	-	-	85.52	**83.96**	**85.20**	87.49	-	-
	20%	20.06	20.53	20.74	20.38	-	-	50.76	**49.94**	**47.00**	50.26	-	-
Thyroid	2%	2.00	2.09	1.80	2.06	-	-	16.91	**14.72**	15.80	**14.32**	-	-
	5%	5.64	6.06	4.80	5.10	-	-	6.31	**5.65**	6.63	**6.07**	-	-
	20%	21.41	21.41	20.28	21.64	-	-	**0.88**	1.37	**0.40**	0.68	-	-
Mushroom	2%	2.00	2.27	1.63	1.24	1.76	**40.49**	53.47	**35.40**	**20.32**	49.73	0.03	**0.01**
	5%	5.47	5.97	4.20	7.59	4.88	**57.18**	**6.47**	12.22	**0.19**	7.86	**0.00**	0.01
	20%	21.54	**28.22**	20.60	**50.51**	20.02	**57.18**	**0.01**	0.29	0.01	**0.00**	**0.00**	0.01
Sonar	2%	1.85	1.87	1.91	1.91	-	-	**98.75**	99.14	**99.18**	99.23	-	-
	5%	6.10	6.00	4.54	5.64	-	-	**90.39**	93.68	**89.05**	90.80	-	-
	20%	21.49	21.98	18.82	22.18	-	-	**63.03**	69.90	61.60	**61.36**	-	-
USPS	2%	1.86	1.66	1.90	1.66	1.87	1.95	**45.43**	62.79	**34.18**	54.38	**19.00**	39.29
	5%	5.24	5.30	4.75	4.71	4.90	5.10	**13.02**	25.51	**11.94**	18.99	**3.60**	8.30
	20%	20.20	21.06	19.95	20.33	19.88	20.11	**0.32**	2.33	**0.10**	1.15	**0.02**	0.40

4 Conclusion

Nous proposons une méthode non paramétrique pour détecter si un point testé est une anomalie par rapport à un comportement nominal. On se sert des propriétés de la décomposition en vecteurs propres de la matrice de Gram d'un noyau $\boldsymbol{K}$.Plus précisément, la statistique de test proposée ici est basée sur le fait que la projection de points nominaux possède une erreur de reconstruction faible dans l'espace engendré par l'ACP à noyau, et que les anomalies sont projetées avec une plus grande erreur de reconstruction.

Cette approche permet notamment de s'adapter au cadre de la grande dimension, ce qui demeure un challenge en détection d'anomalies. Non paramétrique, la méthode

ne fait pas d'hypothèse sur la distribution nominale, et il n'y a pas de paramètres à déterminer. Enfin, comparée à d'autres méthodes usuelles pour la détection d'anomalies sur des données artificielles et réelles, cette approche s'avère compétitive.

Références

[1] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection : A survey. *ACM Comput. Surv.*, 41(3) :15 :1–15 :58, July 2009.

[2] C.-C. Chang and C.-J. Lin. LIBSVM : A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2 :27, 2011.

[3] L. Chapel and C. Friguet. Anomaly detection with score functions based on the reconstruction error of the kernel PCA. In *European Conference on Machine Learning (ECML PKDD)*, volume 8724, pages 227–241, 2014.

[4] H. Hoffmann. Kernel PCA for novelty detection. *Pattern Recognition*, 40(3) :863 – 874, 2007.

[5] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12 :181–201, 2001.

[6] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13 :1443–1471, 2001.

[7] John Shawe-Taylor, Christopher KI Williams, Nello Cristianini, and Jaz Kandola. On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA. *Information Theory, IEEE Transactions on*, 51(7) :2510–2522, 2005.

[8] C. Williams and M. Seeger. The effect of the input density distribution on kernel-based classifiers. In *ICML 17*, pages 1159–1166, 2000.

[9] M. Zhao and V. Saligrama. Anomaly detection with score functions based on nearest neighbor graphs. In *NIPS 22*, pages 2250–2258, 2009.