

# Mesures d'indépendance dans des rkHs en limite plate

Pierre-Olivier AMBLARD<sup>1</sup>, Simon BARTHELME<sup>1</sup>, Nicolas TREMBLAY<sup>1</sup>, Konstantin USEVICH<sup>2</sup>

<sup>1</sup>GIPSAIab/GAIA, CNRS, Univ. Grenoble Alpes, Saint Martin d'Hères, France

<sup>2</sup>CRAN, CNRS, Univ. Lorraine, Nancy, France

pierre-olivier.amblard, simon.barthelme, nicolas.tremblay, konstantin.usevich@cnrs.fr

**Résumé** – Nous examinons les mesures d'indépendance fondées sur les normes de Hilbert-Schmidt d'opérateurs de corrélation entre espaces de Hilbert à noyaux reproduisants dans la limite où les paramètres d'échelle des noyaux tendent vers l'infini. Cette asymptotique appelée limite plate est tout sauf idiote et conduit à des limites non triviales pour certaines machines à noyau. Utilisant des résultats récents sur la limite plate de la régression par processus Gaussiens, nous formalisons la limite pour les mesures d'indépendance HSIC, et illustrons dans quelques cas particuliers et sur des exemples synthétiques le potentiel de ces limites.

**Abstract** – The goal of the paper is a study of the flat limit of reproducing kernel Hilbert spaces (rkHs) based independence measures. We focus on HSIC, the Hilbert-Schmidt Independence Criterion, which consists in measuring the norm of the correlation operator generated by two random variables when embedded in two rkHs. The flat limit examines the asymptotics when the scale parameters of the kernels go to infinity. We use recent results on the flat limit for regression using Gaussian processes to formalize the flat limit of HSIC. We show the potential of the limits for some particular cases and synthetic examples.

**Mesurer l'indépendance ?** Décider de l'indépendance entre deux variables aléatoires  $X$  et  $Y$  met en jeu des mesures de distance ou de divergence entre lois de probabilités, et conduit à des problèmes d'estimation difficiles. Une alternative consiste à utiliser la covariance  $\text{Cov}[f(X), g(Y)]$  entre transformées non-linéaires des variables. Un théorème de Rényi montre que  $\text{Cov}[f(X), g(Y)] = 0$  pour toutes fonctions  $f$  et  $g$  continues bornées implique l'indépendance [11]. En pratique, vérifier le théorème de Rényi est évidemment impossible. Une première possibilité est de vérifier l'annulation des moments croisés de type  $\text{Cov}[X^n, Y^m]$  jusqu'à un certain ordre. Cette condition intuitive, utilisée par exemple dans les premiers algorithmes de séparation de sources indépendantes [10], est évidemment nécessaire, mais non suffisante. Une deuxième possibilité est d'utiliser des fonctions  $f$  et  $g$  dans des espaces de Hilbert à noyaux reproduisants (rkHs, reproducing kernel Hilbert space) bien choisis, et conduit à des conditions suffisantes.

Ce papier montre que l'approche fondée sur les rkHs se simplifie par une asymptotique particulière en l'approche fondée sur l'annulation des moments croisés.

**Mesures d'indépendance dans des espaces de Hilbert à noyau reproduisant.** L'intuition donnée plus haut est formalisée par la généralisation du théorème de Rényi à des espaces de Hilbert à noyaux reproduisants. Si les espaces sont suffisamment riches, mesurer l'indépendance peut s'effectuer en s'assurant que les variables aléatoires à valeurs dans  $\mathbb{R}$   $X$  et  $Y$  sont décorrélées par application de toutes fonctions des espaces. Pour être plus précis, si  $\mathcal{H}_x$  et  $\mathcal{H}_y$  sont deux rkHs associés aux noyaux (caractéristiques)  $k_x$  et  $k_y$  de  $\mathbb{R} \times \mathbb{R}$  dans  $\mathbb{R}$ , alors  $X$  et  $Y$  sont indépendantes si et seulement si  $\sup_{f \in \mathcal{U}_x, g \in \mathcal{U}_y} \text{Cov}[f(X), g(Y)] = 0$ ,  $\mathcal{U}$  désignant la boule unité de  $\mathcal{H}$  [9].

Considérons l'opérateur de covariance  $\Sigma_{XY}$ , opérateur linéaire de  $\mathcal{H}_y$  dans  $\mathcal{H}_x$  défini par  $\langle f | \Sigma_{XY} g \rangle_x = \text{Cov}[f(X), g(Y)]$ , où  $\langle \cdot | \cdot \rangle_z$  dénote le produit scalaire dans  $\mathcal{H}_z$ . Alors

$\sup_{f \in \mathcal{U}_x, g \in \mathcal{U}_y} \text{Cov}[f(X), g(Y)] = \|\Sigma_{XY}\|$ . Comme la norme d'opérateur est majorée par la norme de Hilbert-Schmidt, et comme l'opérateur de covariance est un opérateur de Hilbert-Schmidt, l'équivalence précédente peut se réécrire  $X$  et  $Y$  sont indépendantes si et seulement si  $\|\Sigma_{XY}\|_{HS} = 0$ . En dimension finie, cette norme est la trace de  $\Sigma \Sigma^\top$ . La mesure  $\|\Sigma_{XY}\|_{HS}^2$  est appelée HSIC—pour Hilbert-Schmidt Independence Criterion— et peut facilement être estimée à partir de  $N$  réalisations des variables  $X$  et  $Y$  [9, 7]. On notera  $\widetilde{\mathbf{K}}_X$  la matrice de Gram centrée, d'éléments  $\widetilde{\mathbf{K}}_{Xij} = \langle k(\cdot, X_i) - N^{-1} \sum_l k(\cdot, X_l) | k(\cdot, X_j) - N^{-1} \sum_l k(\cdot, X_l) \rangle_x$ . Si  $\mathbf{C}_N = \mathbf{I}_N - N^{-1} \mathbf{1}\mathbf{1}^\top$ , alors  $\widetilde{\mathbf{K}}_X = \mathbf{C}_N \mathbf{K}_X \mathbf{C}_N$ ,  $\mathbf{K}_X$  étant la matrice de Gram non centrée.  $\Sigma_{XY}$  peut être estimé par  $\widehat{\Sigma}_{XY} = N^{-1} \mathbf{K}_X \mathbf{C}_N \mathbf{K}_Y$ , et sa norme de Hilbert-Schmidt par  $\|\widehat{\Sigma}_{XY}\|_{HS}^2 = N^{-2} \text{Tr}[\widetilde{\mathbf{K}}_X \widetilde{\mathbf{K}}_Y]$ .

Une mesure analogue mais normalisée est donnée par la norme de l'opérateur de corrélation ou covariance normalisée. Cet opérateur est l'unique opérateur  $V_{XY}$  vérifiant  $\Sigma_X^{1/2} V_{XY} \Sigma_Y^{1/2} = \Sigma_{XY}$ . L'écriture prend en compte le fait que l'opérateur de covariance  $\Sigma_X$  n'est pas nécessairement inversible. On abusera la notation en écrivant  $V_{XY} = \Sigma_X^{-1/2} \Sigma_{XY} \Sigma_Y^{-1/2}$  en se rappelant que les inverses des racines carrées des opérateurs de variance ne sont à considérer que sur les espaces engendrés par les fonctions propres associées à des valeurs propres non nulles. Moyennant cette subtilité, la norme de Hilbert-Schmidt de  $V_{XY}$  s'estime à partir de  $N$  réalisations des variables  $X$  et  $Y$  par  $\|\widehat{V}_{XY}\|_{HS}^2 = \text{Tr}[\widetilde{\mathbf{K}}_X \widetilde{\mathbf{K}}_X^{\text{reg}-1} \widetilde{\mathbf{K}}_Y \widetilde{\mathbf{K}}_Y^{\text{reg}-1}]$ , où  $\widetilde{\mathbf{K}}_X^{\text{reg}-1} = (\widetilde{\mathbf{K}}_X + N \lambda \mathbf{I}_N)^{-1}$  est une inverse régularisée de  $\widetilde{\mathbf{K}}_X$ .

Fukumizu a montré que la norme de Hilbert-Schmidt de l'opérateur de corrélation est indépendant du noyau choisi et n'est autre que la contingence [7], soit

$$\|\widehat{V}_{XY}\|_{HS}^2 = \int (p_{X,Y}(x, y) - p_X(x)p_Y(y))^2 dx dy$$

Estimer la contingence peut alors s'effectuer en utilisant  $\|\widehat{V}_{XY}\|_{HS}^2$

dont la convergence asymptotique a également été montré par Fukumizu (moyennant que  $\lambda$  tende vers zero avec  $N$ , mais pas trop vite).

**Limite plate.** Les noyaux considérés ici sont radiaux, et s'écrivent  $k(\mathbf{x}, \mathbf{y}) = \gamma f(\varepsilon \|\mathbf{x} - \mathbf{y}\|)$  où  $f$  est une fonction définie-positve. La limite plate consiste à examiner le comportement de machines à noyaux (régression, classifieurs, ...) lorsque le paramètre d'échelle  $\varepsilon^{-1}$  tend vers l'infini. Cette limite peut paraître idiote à étudier puisque les matrices de Gram associées deviennent singulières (les éléments tendent vers une même constante). Toutefois, comme initié dès 2002 par Driscoll&Fornberg dans le cas de l'interpolation [5], et comme étudié profondément récemment par Barthelmé&Usvich pour les noyaux en général [2], la dynamique en  $\varepsilon$  avec laquelle ces matrices se singularisent permet l'existence de limites non triviales pour certaines machines à noyaux. Nous avons par exemple montré que les processus déterminantaux en limite plate restent bien définis [3], exhibé les liens profonds en limite plate entre régression par processus gaussiens et splines polyharmoniques [4]. Dans ce dernier exemple en particulier, il s'avère, bien que ce ne soit pas démontré, que dans beaucoup de cas pratiques, la limite est atteinte pour des gammes du paramètre  $\varepsilon$ , justifiant l'utilisation de la limite à la place de la machine à noyau, éliminant *de facto* la calibration de paramètres.

Dans cette contribution, nous nous posons la question de l'intérêt de la limite plate pour les mesures d'indépendance fondées sur des noyaux [9, 7, 8]. Ce travail préliminaire apporte quelques idées.

**HSIC et limite plate.** Dans l'asymptotique étudiée, le nombre d'échantillons permettant l'estimation est fixe et fini, et nous examinons le comportement de l'estimateur de HSIC utilisant la corrélation lorsque les paramètres d'échelle des noyaux en jeu tendent vers l'infini. Un problème très proche a été étudié récemment par Alvarez [1], dans le cadre de l'alignement de noyau. La mesure étudiée par Alvarez est HSIC normalisée *a posteriori* (c'est-à-dire la norme de l'opérateur de covariance, normalisée par les normes des opérateurs de variance). Dans son travail, seul le noyau Gaussien est utilisé, et il ne considère que le premier ordre dans les perturbations.

L'approche ici est plus générale. Les rkHs dans lesquels sont plongés les deux variables aléatoires peuvent être différents. Les noyaux sont stationnaires et radiaux  $k(\mathbf{x}, \mathbf{y}) = \gamma f(\varepsilon \|\mathbf{x} - \mathbf{y}\|)$ . Ils incluent comme cas particulier le noyau Gaussien, mais englobent également des noyaux plus généraux, moins réguliers, comme les noyaux de Matérn par exemple. La régularité  $r$  des noyaux a par ailleurs une grande importance dans les résultats que nous apportons.  $r$  est défini comme le premier degré pour lequel le noyau n'est pas dérivable. Par exemple pour le gaussien  $r = \infty$ , pour le noyau exponentiel  $k(u) = \exp(-|u|)$   $r = 1$ , pour  $k(u) = (1 + |u|) \exp(-|u|)$   $r = 2, \dots$  Typiquement, un noyau de régularité  $r$  admet comme développement de Taylor une série du type  $k(u) = f_0 + f_2|u|^2 + \dots + f_{2r-2}|u|^{2r-2} + f_{2r-1}|u|^{2r-1} + \dots$  et  $r$  quantifie alors l'apparition du premier degré impair dans ce développement.

Nous ne présentons ici les résultats que pour des variables aléatoires réelles. L'extension aux vecteurs aléatoires, si elle ne pose pas de problèmes conceptuels supplémentaires, nécessite des résultats sur les polynômes multivariés plus complexes à présenter et qui peuvent troubler le message. Mais les idées relatées ici restent valables dans un cadre plus général.

## 1 Quelques rappels de résultats récents

Reprenons l'estimateur de HSIC (on entend dans toute la suite la version reposant sur la corrélation),  $\|\widehat{V}_{XY}\|_{HS}^2 = \text{Tr}[\widetilde{\mathbf{K}}_X \widetilde{\mathbf{K}}_X^{reg-1} \widetilde{\mathbf{K}}_Y \widetilde{\mathbf{K}}_Y^{reg-1}]$ . Les matrices relatives à une des deux

variables, e.g.  $\mathbf{M}_X = \widetilde{\mathbf{K}}_X \widetilde{\mathbf{K}}_X^{reg-1}$  intervenant dans la définition ne sont rien d'autre que des matrices de lissage associé à des régressions à noyaux.

Pour être précis, supposons que l'on observe  $N$  mesures  $z_i$  modélisées par  $z_i = f(x_i) + w_i$ , où les  $x_i$  sont connus, la suite  $w_i$  une suite de variables aléatoires i.i.d. centrées, normales de variance  $\sigma^2$ , et que l'on souhaite inférer  $f(x)$ . La solution donnée par une approche bayésienne ayant comme *a priori* pour  $f$  un processus gaussien centré et de covariance  $k(x, y)$  est résumée par  $E[f(x)|\mathbf{z}] = \mathbf{k}_{x,\mathbf{x}}(\mathbf{K}\mathbf{x} + \sigma^2\mathbf{I})^{-1}\mathbf{z}$  où  $\mathbf{k}_{x,\mathbf{x}} = [k(x, x_1), \dots, k(x, x_N)]$ . Cette relation valable pour tout  $x$  permet donc d'obtenir l'estimateur des  $f(x_i)$ , autrement dit le lissage de  $\mathbf{z}$  par la relation  $\mathbf{K}\mathbf{x}(\mathbf{K}\mathbf{x} + \sigma^2\mathbf{I})^{-1}\mathbf{z}$ . On appelle ainsi  $\mathbf{M}\mathbf{x} = \mathbf{K}\mathbf{x}(\mathbf{K}\mathbf{x} + \sigma^2\mathbf{I})^{-1}$  la matrice de lissage. Son importance est capitale dans l'analyse de la machine à noyau associée. On montre par exemple que la variance de prédiction  $\text{Var}[f(x)|\mathbf{z}]$  peut s'écrire uniquement en fonction de  $\sigma^2$  et de  $\mathbf{M}_{(x,\mathbf{x})}$  [4]; par ailleurs,  $\text{Tr}\mathbf{M}\mathbf{x}$  est une mesure naturelle du nombre de degrés de liberté de la machine à noyau [6].

Avant d'énoncer le résultat important, introduisons quelques notations. On note  $\mathbf{V}_x$  la matrice de Vandermonde associée aux données  $x_1, \dots, x_N$ , c'est-à-dire la matrice des monômes jusqu'à l'ordre  $N - 1$  de ces données. La  $i$ ème colonne s'écrit  $\mathbf{v}_i = (x_1^i, \dots, x_N^i)^\top$  où  $i = 0, 1, \dots, N - 1$ . On note  $\mathbf{V}_{x, \leq r}$  la restriction de  $\mathbf{V}_x$  à ses  $r$  premières colonnes (ou  $< r$  pour les  $r - 1$  premières). Lorsqu'il n'y a pas ambiguïté, on supprime l'indice  $x$  dans les notations. Soit  $\mathbf{D}^{(2r-1)} = [ |x_i - x_j|^{2r-1} ]_{i,j}$  la matrice des distances élevées à la puissance  $2r - 1$ . On définit

$$\widetilde{\mathbf{D}}^{(2r-1)} = (\mathbf{I} - \mathbf{Q}_{\leq r} \mathbf{Q}_{\leq r}^\top) \mathbf{D}^{(2r-1)} (\mathbf{I} - \mathbf{Q}_{\leq r} \mathbf{Q}_{\leq r}^\top) \quad (1)$$

où  $\mathbf{Q}_{< r}$  est une base orthogonale pour l'espace engendré par les colonnes de  $\mathbf{V}_{< r}$ , obtenue typiquement en utilisant la décomposition QR de  $\mathbf{V}_x = \mathbf{Q}\mathbf{R}$ .  $\widetilde{\mathbf{D}}^{(2r-1)}$  correspond à  $\mathbf{D}^{(2r-1)}$  dans laquelle les monômes de degrés inférieurs à  $r$  sont éliminés par projections sur les sous-espace orthogonaux à  $\mathbf{V}_{< r}$ .

En utilisant le comportement spectral en limite plate des matrices de Gram associées à des noyaux stationnaires [2], nous montrons dans [4] le résultat suivant ( $\tilde{\gamma}$  est un réel fixé) :

**Théorème 0.1** *Soit  $k(\varepsilon|x - y|)$  un noyau défini-positif stationnaire sur  $\mathbb{R}$ , de régularité  $r$ . Soit  $p \in \mathbb{N}$ . Alors la matrice de lissage  $\mathbf{M}_\varepsilon = \tilde{\gamma} \varepsilon^{-p} \mathbf{K}(\tilde{\gamma} \varepsilon^{-p} \mathbf{K} + \sigma^2 \mathbf{I})^{-1}$  associée à  $\tilde{\gamma} \varepsilon^{-p} k_\varepsilon$  converge lorsque  $\varepsilon \rightarrow 0$  vers :*

- pour  $p > 2r - 1$ , l'identité (interpolation)
- pour  $p = 2r - 1$ ,  $\mathbf{Q}_{\leq r} \mathbf{Q}_{\leq r}^\top + \widetilde{\mathbf{D}}^{(2r-1)} (\widetilde{\mathbf{D}}^{(2r-1)} + \frac{\sigma^2}{\tilde{\gamma}} \mathbf{I})^{-1}$ . Cette matrice limite correspond à la matrice de lissage de fonction splines d'ordre  $r$  pénalisée et ajoutée à la régression polynomiale d'ordre  $r - 1$ .
- pour  $p < 2r - 1$  impair,  $\mathbf{Q}_{\leq l} \mathbf{Q}_{\leq l}^\top$ ,  $l = \lfloor p/2 + 1 \rfloor$ , qui correspond à une régression polynomiale d'ordre  $l - 1$ .
- pour  $p < 2r - 1$  pair,  $\mathbf{Q}_{\leq l} \mathbf{Q}_{\leq l}^\top + \frac{\tilde{\lambda}_{p/2+1}}{\tilde{\lambda}_{p/2+1} + \frac{\sigma^2}{\tilde{\gamma}}} \mathbf{q}_{p/2+1} \mathbf{q}_{p/2+1}^\top$ ,  $l = p/2$  qui correspond à une régression polynomiale d'ordre  $l$ , où seul le dernier monôme est pénalisé.

Les  $\tilde{\lambda}_i$  apparaissant dans le théorème correspondent au comportement dominant de la  $i$ ème valeur propre de  $\mathbf{K}$ . On montre en effet [2] que  $\lambda_i(\varepsilon) = \varepsilon^{\alpha_i} (\tilde{\lambda}_i + O(\varepsilon))$ , où  $\alpha_i = 2i$  pour  $i = 0, \dots, r - 1$ , et  $\alpha_i = 2r - 1$  pour  $i = r, \dots, N - 1$ .

On note dans ce résultat le caractère fondamental de la renormalisation en amplitude du noyau par  $\gamma := \tilde{\gamma} \varepsilon^{-p}$  pour que la limite plate ait

un sens intéressant. En son absence, on ne conserve dans les résultats précédents que la constante !

## 2 HSIC en limite plate

L'utilisation des résultats précédents permet d'obtenir facilement le comportement de HSIC dans l'asymptotique de la limite plate. Dans la suite, les matrices de Gram considérées sont renormalisées en amplitude par  $\gamma = \tilde{\gamma}\varepsilon^{-p}$ . *A priori*, le noyau  $k, \tilde{\gamma}$  et  $p$  dépendent de la variable ( $X$  ou  $Y$ ). Nous ne ferons apparaître les indices  $X$  ou  $Y$  qu'en cas de besoin.

Deux différences minimales sont à considérer. Les matrices  $M_X = \widetilde{K_X} K_X^{reg-1}$  et  $M_Y$  sont des matrices de lissage construites sur les noyaux centrés, et il convient de remplacer  $\sigma^2$  par une régularisation dépendant de  $N$ , soit  $N\sigma^2$ . Ce choix est effectivement arbitraire mais permet de rester dans le cadre de l'analyse asymptotique de Fukumizu en ce qui concerne la convergence des estimateurs des opérateurs de covariance. Le centrage revient à éliminer le premier vecteur propre de  $M$ . En effet, à  $\varepsilon$  fixe, comme  $M_X = \gamma \widetilde{K_X} (\gamma \widetilde{K_X} + N\sigma^2 \mathbf{I}_N)^{-1}$ , elle admet le développement sur la base des vecteurs propres sans  $\mathbf{u}_0(\varepsilon)$  de  $K_X$  selon

$$M_X(\varepsilon) = \sum_{i=1}^{N-1} \frac{\gamma \lambda_i(\varepsilon)}{\gamma \lambda_i(\varepsilon) + N\sigma^2} \mathbf{u}_i(\varepsilon) \mathbf{u}_i(\varepsilon)^\top$$

Les  $\lambda_i(\varepsilon)$  et  $\mathbf{u}_i(\varepsilon)$  sont les éléments propres de  $K_X$  (et seront indicés par  $X$  si nécessaire), et les résultats du théorème 1 restent valides, à condition de ne pas prendre en compte la première colonne de  $\mathbf{Q}$ .

Avant d'appliquer le théorème, il est intéressant d'expliciter HSIC à  $\varepsilon$  fixe. En utilisant la représentation spectrale précédente, on obtient

$$\begin{aligned} \|\widehat{V_{XY}}(\varepsilon)\|_{HS}^2 &= \text{Tr}[M_X(\varepsilon)M_Y(\varepsilon)] \\ &= \sum_{i,j=1}^{N-1} \frac{\gamma_X \gamma_Y \lambda_{X,i}(\varepsilon) \lambda_{Y,i}(\varepsilon) [\mathbf{u}_i^X(\varepsilon)^\top \mathbf{u}_j^Y(\varepsilon)]^2}{(\gamma_X \lambda_{X,i}(\varepsilon) + N\sigma^2)(\gamma_Y \lambda_{Y,i}(\varepsilon) + N\sigma^2)} \end{aligned}$$

qui montre la décomposition de HSIC en somme pondérée des produits scalaires entre vecteurs propres des matrices de Gram. En anticipant un peu la discussion qui suit, cette remarque est intéressante puisque les limites des vecteurs propres de  $K_X$  sont intimement liés aux polynômes orthogonaux des mesures empiriques construites sur les données. Les produits scalaires à la limite apparaissent alors comme somme de moments d'ordres supérieurs croisés empiriques. A titre d'exemple, on illustre les premiers vecteurs propres de  $K_X$  et  $K_Y$  sur un exemple dans la figure 1, où l'on voit clairement des comportements polynômiaux d'ordres croissants.

Si on considère maintenant que les paramètres  $p$  et les régularités des deux noyaux peuvent dépendre de  $X$  et  $Y$ , la mesure HSIC peut prendre 9 formes possibles lorsque les deux noyaux ont une régularité finie (en supposant que les cas triviaux  $p > 2r + 1$  sont exclus), 6 si l'un des deux noyaux a une régularité infinie, et 4 si les deux sont infiniment réguliers.

Nous considérons deux situations en particulier :

1.  $r_X$  fini (3 pour les illustrations) et  $r_Y$  infini;  $p_X = 2r_X - 1$  et  $p_Y$  impair. Dans ce cas, les matrices limites sont

$$M_X = \mathbf{Q}_{X,2 \leq r_X} \mathbf{Q}_{X,2 \leq r_X}^\top + \widetilde{D}^{(2r_X-1)} (\widetilde{D}^{(2r_X-1)} + \frac{N\sigma^2}{\tilde{\gamma}} \mathbf{I})^{-1}$$

$$M_Y = \mathbf{Q}_{Y,2 \leq l_Y} \mathbf{Q}_{Y,2 \leq l_Y}^\top$$

$$\text{où } l_Y = \lfloor p_Y/2 + 1 \rfloor$$

2.  $r_X$  et  $r_Y$  infinis;  $p_X$  et  $p_Y$  impairs. Alors  $M_Y$  est définie comme précédemment, et  $M_X = \mathbf{Q}_{X,2 \leq l_X} \mathbf{Q}_{X,2 \leq l_X}^\top$

Alors  $\text{Tr}(M_X M_Y)$  est la limite plate de  $\|\widehat{V_{XY}}\|_{HS}^2$  pour les paramètres donnés.

## 3 Illustrations et discussions.

Pour les illustrations qui suivent, nous mesurons l'indépendance entre  $X$  et  $Y = \alpha X + W$  où  $W$  est une variable indépendante de  $X$ , choisie normale centrée réduite. Nous faisons varier  $\alpha$  de 0 à 2, et représentons les différentes mesures en fonction de  $\alpha$ .  $X$  suit une loi exponentielle de moyenne 2, et on utilise ici 100 réalisations i.i.d. de ces variables.

La première illustration (Fig. 1) concerne la limite plate calculée numériquement en utilisant les noyaux. Pour montrer la convergence en limite plate de HSIC, on fixe un nombre de degré de liberté  $l = \text{Tr}(M)$  pour les machines à noyaux (ici  $l = 4$ ), équation que l'on résoud pour  $\gamma$  en fonction de  $\varepsilon$ . Les courbes bleues dans les tracés droits de la figure 1 correspondent au calcul de HSIC pour  $\varepsilon$  décroissant, et  $\gamma$  choisi pour que  $l = 4$ . Ces courbes convergent vers la limite plate représentée en pointillés rouges. Le tracé fonction de  $\alpha$  donne sur cet exemple très simple une idée de la sensibilité à l'indépendance de la mesure. Les deux tracés à gauche et au centre montre les premiers vecteurs propres de  $K_X$  et  $K_Y$  (et donc ceux des matrices de lissages associées) qui interviennent dans le calcul de HSIC et correspondant à la limite plate. Les noyaux identiques pour  $X$  et  $Y$  ont pour régularité 3 pour la ligne inférieure (noyau  $k(u) = (1 + |u| + |u|^2/3) \exp(-|u|)$ ), et  $\infty$  pour la ligne supérieure (noyau gaussien).

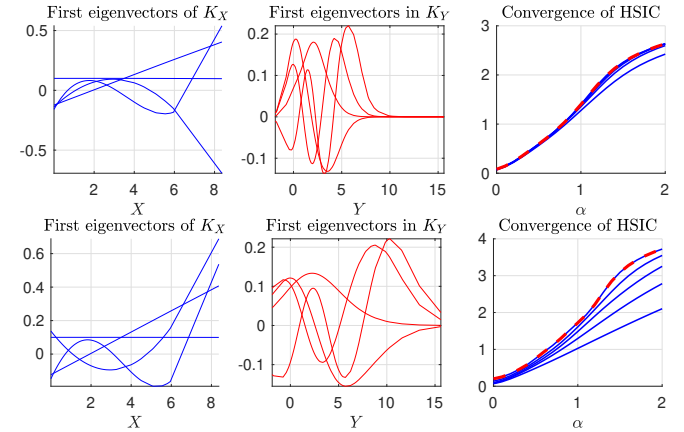


FIGURE 1 – Haut :  $r_X = r_Y = \infty$ . Bas  $r_X = r_Y = 3$ . Gauche et milieu : tracés des quatre premiers vecteurs propres utilisés. Droite, convergence de HSIC vers sa limite plate (pointillés rouges), calculée pour un nombre de degré de liberté fixé à 4.  $\alpha = 0$  correspond à une situation d'indépendance.

Dans un deuxième temps, on considère les estimateurs construits directement à partir des équations donnant les limites plates, dans les deux situations 1- et 2- développées précédemment. Le problème analysé est le même que dans la première illustration. L'objet n'est pas ici de faire une analyse statistique de ces mesures, mais pour montrer la dispersion, nous traçons 100 réalisations de l'expérience dans les deux cas, et pour deux valeurs  $N = 100$  et  $500$  du nombre d'observations indépendantes des variables  $X$  et  $Y$ . De plus, pour donner un aperçu du caractère discriminant de ces mesures limites, nous traçons en noir le résultat des calculs lorsque les  $N$  observations  $X$  sont per-

mutées aléatoirement, de sorte à simuler l'indépendance pour tout  $\alpha$ . Les paramètres sont donnés dans la légende de la figure 2.

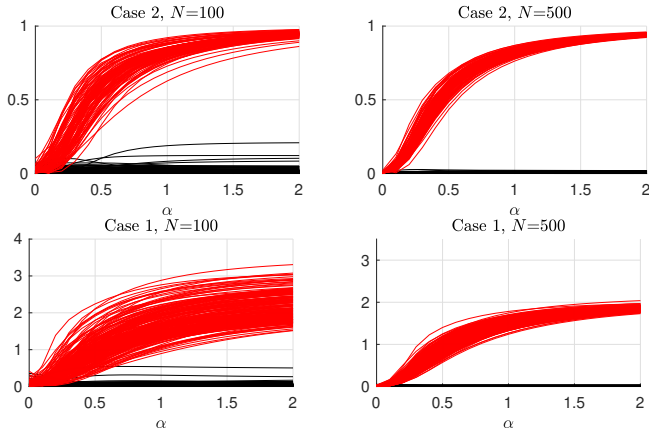


FIGURE 2 – Utilisation des limites. Superposition de 100 réalisations du calcul pour rendre compte de la dispersion. Courbe rouge : calcul sur les données ; courbes noires : calcul sur les données  $x_i$  permutées aléatoirement. Haut :  $r_X = r_Y = \infty, p_X = 3$  et  $p_Y = 5$ . Bas  $r_X = 3, r_Y = \infty, p_X = 2r_X - 1$  et  $p_Y = 5$ . Gauche : pour  $N = 100$  observations. Droite : pour  $N = 500$  observations

Il est remarquable de constater que ces mesures limites semblent avoir un comportement suffisant pour discriminer l'indépendance de la dépendance. Cette assertion ne vaut pas preuve bien évidemment. D'ailleurs, il est peu probable que la nullité de HSIC limite dans ces deux cas conduise à une condition nécessaire et suffisante d'indépendance. En effet, la mesure ne fait intervenir que des moments croisés empiriques d'ordres supérieurs, mais d'ordre fini.

Enfin, on illustre la limite plate de HSIC dans le cas 2,  $\text{Tr} [Q_{X,2 \leq l_X} Q_{X,2 \leq l_X}^\top Q_{Y,2 \leq l_Y} Q_{Y,2 \leq l_Y}^\top]$ , que l'on compare à HSIC (calculé avec le noyau Gaussien, d'écart type l'écart type de la variable aléatoire concernée) pour l'exemple issu de [7]. Ici, le couple  $X, Y$  est la rotation d'angle  $\alpha$  d'un couple de variables indépendantes (Uniforme sur  $[0, 2]$  ; Uniforme sur  $[-0.75, -0.25] \cup [0.25, 0.75]$ ).  $X$  et  $Y$  sont donc indépendantes pour les rotations d'angle multiples de  $\pi/2$ . Deux réalisations pour des angles différents sont présentées figure 3. On représente les résultats pour HSIC et sa limite plate, pour deux tailles d'échantillons  $N = 100$  et  $500$ , et pour 100 réalisations indépendantes. Comme précédemment, le calcul pour les échantillons  $X$  permutés aléatoirement est tracé, de sorte à simuler l'indépendance. On superpose sur ces graphes la moyenne des mesures (en jaune). La limite plate a un comportement très satisfaisant, puisque l'indépendance est détectée. De manière non suprenante, les statistiques pour la limite plate sont sensiblement meilleures que celles d'HSIC. Dans le cas présent, les polynômes en jeu dans la limite plate sont linéaires en  $X$  et cubiques en  $Y$ . Les statistiques croisées à estimer sont donc d'ordre 4, contrairement à HSIC pour lequel l'ensemble des ordres est sous-jacent. Par ailleurs, aucun paramètre n'est à régler dans la mesure limite plate.

Ces premières idées doivent maintenant être creusées. Plusieurs pistes doivent être suivies. Pratiquement, une analyse statistique doit être menée. Théoriquement, des questions se posent quant à la validité des limites plates pour théoriquement mesurer l'indépendance. De manière intéressante, les mesures obtenues généralisent les mesures construites à partir de noyaux définis-positifs à des mesures bâties sur des noyaux conditionnellement définis-positifs, pour les-

quelles la théorie de opérateurs de covariance n'est pas disponible à notre connaissance.

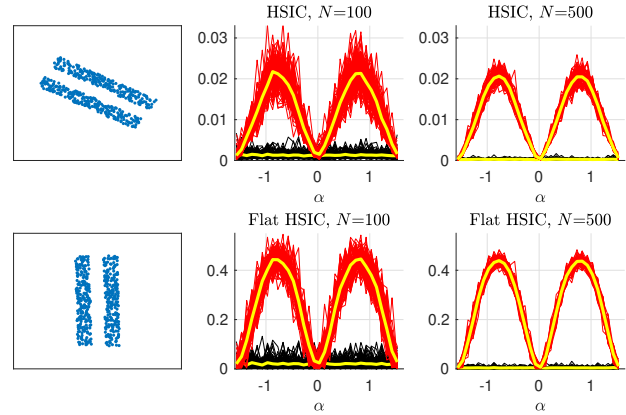


FIGURE 3 – Limite plate de HSIC (cas 2) contre HSIC pour un deuxième exemple. Le couple  $X, Y$  est la rotation d'angle  $\alpha$  d'un couple  $(U([0, 2]); U([-0.75, -0.25] \cup [0.25, 0.75]))$ . Les variables sont indépendantes aux multiples de  $\alpha = \pi/2$ . On représente colonne de gauche deux réalisations pour deux angles différents. la colonne du milieu est le résultat pour  $N = 100$  échantillons, la colonne de droite pour  $N = 500$ . HSIC (ligne sup.) est calculée avec un noyau gaussien, d'écart type la déviation standard des données. La limite plate (cas 2) est calculée avec  $p_X = 3$  et  $p_Y = 5$  (polynômes linéaires en  $X$ , cubiques en  $Y$ ). En rouge 100 réalisations pour chaque cas, en noir 100 réalisations avec  $X$  permuté aléatoirement (simulation d'indépendance à partir des données), en jaune les mesures moyennes.

## Références

- [1] S. Alvarez. Gaussian RBF centered kernel alignment (CKA) in the large bandwidth limit. In *arXiv :2112.09305*, 2021.
- [2] S. Barthelmé and K. Usevich. Spectral properties of kernel matrices in the flat limit. *SIAM Journal on Matrix Analysis and Applications*, 42 :17–57, 2020.
- [3] S. Barthelmé, N. Tremblay, K. Usevich, and P.-O. Amblard. Determinantal point processes in the flat limit. *arXiv preprint arXiv :2107.07213*, to appear, Bernoulli, 2022.
- [4] S. Barthelmé, P. O. Amblard, N. Tremblay et K. Usevich, Gaussian Process Regression in the flat limit. ArXiv : arXiv :2201.01074.
- [5] T. A. Driscoll and B. Fornberg. Interpolation in the limit of increasingly flat radial basis functions. *Computers & Mathematics with Applications*, 43(3-5) :413–422, 2002.
- [6] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Springer, 2001.
- [7] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *NIPS*, 2007.
- [8] K. Fukumizu, F. R. Bach, and M. I. Jordan. Kernel dimension reduction in regression. *The Annals of Statistics*, 37(4) :1871–1905, 2009.
- [9] A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6 :2075–2129, 2005.
- [10] C. Jutten and J. Héault. Blind separation of sources, part i : An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24 :1–10, 1991.
- [11] A. Rényi. On measures of dependence. *Acta Math. Acad. Sci. Hungar.*, 10(2) :441–451, september 1959.
- [12] N. Tremblay, S. Barthelmé, K. Usevich, and P.-O. Amblard. Extended  $L$ -ensembles : a new representation for determinantal point processes. *arXiv :2107.06345*, 2021.