# Réduction des coûts algorithmiques par approximation de faible rang pour la détection de ruptures à noyau reproduisant et de grands flux de données (Big Data)

**Projet**

Dans le contexte du machine/statistical learning, les approches basées sur les noyaux reproduisants $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ sont très répandues. Elles présentent entre autres l'avantage de pouvoir traiter des types variés de données (réseaux, séquences d'ADN,...).

Toutefois une limitation de ces approches basées sur les noyaux est qu'elles induisent un coût algorithmique souvent trop important, lié au calcul et au stockage de la matrice de Gram $G = \{k(x_i, x_j)\}_{1 \leq i,j \leq n}$ de taille $n \times n$.

Une approche extrêmement efficace et prometteuse pour lever cette limitation est d'approcher la matrice $G$ par une matrice de faible rang. Plusieurs approches ont été proposées dans la littérature telles que l'approximation de Nyström, l'utilisation des Random Fourier features,... Le point essentiel est que ces techniques sont suffisamment génériques pour être utilisées dans presque tous les cadres de machine learning où les noyaux reproduisants sont utilisés.

Le cadre privilégié du présent stage est celui de la détection de ruptures, qui consiste à détecter automatiquement des plages temporelles où un système d'étude (moteur, réseau social, cours de bourse,...) se comporte de façon atypique par rapport à ce qui précédait et ce qui va suivre. Une procédure de détection de ruptures reposant sur les noyaux reproduisants a été proposée. Cette procédure appelée KCP a déjà montré ses bonnes performances pratiques dans de nombreux contextes différents. Toutefois, son utilisation sur de grands jeux de données ($n \geq 10^6$) reste difficile.

Les objectifs du stage sont :

1. appréhender et mettre en œuvre plusieurs techniques d'approximation de la matrice de Gram par une matrice de faible rang,

2. étudier numériquement (et théoriquement) ces approches afin d'en mesurer la performance et de les comprendre dans le cadre de la détection de ruptures (KCP),

3. sur la base des résultats précédant, développer une nouvelle stratégie pour réduire le temps de calculs et le coût de stockage de KCP (ou d'autres alternatives telles que la segmentation binaire),

4. mettre en œuvre l'approche globale ainsi obtenue sur des données de type biologique et/ou en provenance de bases de données librement disponibles sur le net.

**Prérequis**
Outre de bonnes compétences en probabilités et statistique, le candidat utilisera `R`, Python ou matlab pour réaliser des simulations afin de vérifier empiriquement ses résultats.

*Encadrant :* Alain Celisse, maître de conférences.

*Length:* 4-6 mois.

*Opportunity:* Le stage pourrait donner lieu à une thèse.

*Laboratory:* MODAL équipe-projet Inria, Lille.

*Contact:* Alain Celisse (`celisse@math.univ-lille1.fr`).

*Suite:* Le sujet proposé pourrait faire l'objet d'une thèse.

**Bibliographie**


- Arlot, S., Celisse, A., Harchaoui, Z. (2012). A kernel multiple change-point algorithm via model selection. arXiv preprint arXiv:1202.3878.
- CELISSE, Alain, MAROT, Guillemette, PIERRE-JEAN, Morgane, et al. New efficient algorithms for multiple change-point detection with reproducing kernels. Computational Statistics and Data Analysis, 2018, vol. 128, p. 200-220.
- RUDI, Alessandro et ROSASCO, Lorenzo. Generalization properties of learning with random features. In : Advances in Neural Information Processing Systems. 2017. p. 3215-3225.