Proposition de sujet M2 Maths Appliquées
Estimation non asymptotique de valeurs propres et vecteurs propres de matrices de covariance en grande dimension

Note : ce sujet est particulièrement adapté à un couplage stage d'IMPACT/Mémoire de M2 de mathématiques appliquées.

Il sera co-encadré par Mylène Maïda, Professeur, Université Lille 1, Laboratoire Paul Painlevé et Philippe Preux, Professeur, Université Lille 3, équipe SequeL INRIA-Nord Europe.

Ce sujet de stage est notamment motivé par le fait que les méthodes statistiques (au sens large) doivent faire face aujourd'hui à d'énormes quantités de données de grande dimension et être capable de les traiter efficacement (Big Data). Pour cela, la solution ne passe pas seulement par des ordinateurs plus puissants, mais surtout par des méthodes améliorées, voire renouvelées, et des algorithmes mieux pensés, tirant parti de propriétés mathématiques liées aux données traitées.

Dans les problèmes d'estimation, on est souvent amené à étudier la matrice de covariance empirique associée au problème et en particulier à estimer ses valeurs propres et vecteurs propres. Des estimateurs "classiques", consistants et asymptotiquement gaussiens, sont bien connus dans le cas de données de petite dimension et grande taille d'échantillons. Dans le contexte actuel de données de très grande dimension, ces estimateurs ne sont plus très pertinents. Certains résultats de la théorie des matrices aléatoires permettent de comprendre le comportement de ces matrices de covariance empirique quand la taille de l'échantillon est très grand mais reste comparable à la dimension des données. Dans le stage d'IMPACT, on introduira les outils mathématiques permettant de comprendre certains de ces résultats asymptotiques. On se concentrera ensuite sur l'étude de l'article [1], publié par Mestre en 2008, qui propose un nouvel estimateur, bien adapté au cas où la taille de l'échantillon n'est pas très grande. Le but du stage est à la fois de vérifier par des simulations que l'estimateur proposé par Mestre est meilleur que les autres estimateurs connus et de comprendre la preuve de la consistance de son estimateur. Cette preuve fait appel à la fois aux probabilités et à l'analyse complexe.

Tout au long de ce travail, on mettra en œuvre les résultats théoriques sur des jeux de données réels ; on réalisera pour cela des simulations en *R*. En effet, ce sujet est notamment motivé par des applications pratiques nécessitant la manipulation de données complexes de grande taille (des textes en langue naturelle, des images, des graphes en particulier). De tels jeux de données seront fournis ; des jeux de données artificielles pourront être également conçus à des fins expérimentales. Il sera pertinent d'interpréter les résultats théoriques sur ces données, voire d'étudier l'extension des résultats théoriques à des situations où les hypothèses nécessaires aux démonstrations ne sont pas vérifiées.

Bibliographie : [1] Improved Estimation of Eigenvalues and Eigenvectors of Covariance Matrices Using Their Sample Estimates, Xavier Mestre, IEEE Transactions on Information Theory, Vol 54, no 11, (2008).