SUJET DE MASTER : géométrie de l'information et apprentissage statistique dans les variétés Riemanniennes

ENCADRANTS : Salem Said, Marc Arnaudon, Yannick Berthoumieu

ÉTABLISSEMENT : Laboratoire IMS (bât A31), IMB (bât A33), Université de Bordeaux, Campus de Talence

CV ET LETTRE DE MOTIVATION : envoyer à `salem.said@u-bordeaux.fr`

---

Les problèmes d'apprentissage statistique (par exemple d'estimation de densité de probabilité, de régression non linéaire, ou de classification) sont classiquement posés pour des données qui appartiennent à un espace euclidien. Or, de plus en plus d'applications font appel à des données qui appartiennent à une variété Riemannienne. Par exemple, on voit apparaître de telles données en traitement des signaux radar, en neurosciences, en vision par ordinateur, et en robotique, entre autres. Le sujet de master permettra au candidat de se familiariser avec la recherche menée aux laboratoires IMS et IMB sur l'apprentissage statistique dans les variétés Riemanniennes. C'est une recherche interdisciplinaire, faisant intervenir la science des données et les mathématiques, et qui a déjà produit des résultats bien reconnus.

Deux pistes pourront être poursuivies durant le stage, selon le choix du candidat et des encadrants : (i) l'estimation en ligne des lois de mélange dans une variété Riemannienne, (ii) le calcul de la distance statistique entre deux clusters (populations unimodales) sur une variété Riemannienne. Concrètement, les variétés en question seront des variétés de Grassmann (applications en robotique et en vision par ordinateur, et incontournables pour les problèmes de réduction de dimensionnalité), ou des variétés de matrices de covariance (applications en radar ou en neurosciences).

– Piste (i) : classiquement, l'estimation d'une loi de mélange se fait en utilisant l'algorithme EM (expectation-maximisation). Cet algorithme fait appel à de trop grandes ressources en mémoire lorsqu'il est appliqué à un grand volume de données. Pour éviter ce problème, on propose de mettre en place une version en ligne de l'algorithme EM, qui ne stocke pas toutes les données en mémoire, mais traite plutôt chaque donnée une seule fois. Le candidat devra se familiariser avec les lois de mélange dans les variétés Riemanniennes, et avec l'algorithme EM pour l'estimation de ces lois. L'objectif sera ensuite d'implémenter sous MATLAB ou Python la version en ligne de l'algorithme EM, et de la valider sur des données réelles.

– Piste (ii) : un enjeu important est d'associer à chaque donnée une interprétation concrète. Par exemple, on cherche à associer à un signal radar le type de cible qui l'aurait renvoyé, ou à un signal d'électroencéphalogramme le type de stimulus qui l'aurait déclenché. Pour cela, il faut structurer les bases de données,(acquises au fil des expériences), en clusters, c'est à dire en populations unimodales qui représentent la variabilité des données ayant une interprétation commune. Le calcul d'une distance statistique entre clusters devient alors un

outil important pour la gestion des données.

Le candidat devra se familiariser avec la technique dite de « tir géodésique » qui sera utilisée pour effectuer le calcul de la distance statistique. Cette technique fait appel aux aspects fondamentaux de la géométrie Riemannienne, liés à l'équation de Jacobi. L'objectif sera de l'implémenter sous MATLAB ou Python, et de la valider, à travers l'application à la problématique de la fusion des bases de données.

---

Profil recherché : bonnes connaissances en statistique inférentielle ; la familiarité avec les algorithmes EM, k-moyennes, SVM serait un plus ; quelques connaissances de départ en géométrie différentielle ou Riemannienne ; aisance avec MATLAB ou Python

---

Poursuite : ce stage ouvre potentiellement sur une thèse de doctorat

---

**Bibliographie :**


— Bases en géométrie Riemannienne :
  - Initiation à la géométrie de Riemann. François Rouvière, Calvage et Mounet 2016.
  - Riemannian geometry. Sylvestre Gallot, Dominique Hulin, Jacques Lafontaine, Springer-Verlag 2004

— Données sur les variétés Riemanniennes :
  - Intrinsic statistics on Riemannian manifolds : basic tools for geometric measurements. Xavier Pennec, (https ://hal.inria.fr/inria-00614994).
  - A Riemannian framework for tensor computing. Xavier Pennec, Pierre Fillard, Nicolas Ayache, (https ://hal.inria.fr/inria-00614990).

— Modèles de mélange sur les variétés Riemanniennes :
  - Clustering on the unit hypersphere using von Mises-Fisher distributions. Arindam Banerjee, Inderjit Dhillon, Joydeep Ghosh, Suvrit Sra, (http ://www.jmlr.org/papers/volume6/banerjee05a/banerjee05a.pdf).
  - Riemannian Gaussian distributions on the space of symmetric positive defnite matrices. Salem Said, Lionel Bombrun, Yannick Berthoumieu, Jonathan Manton, (https ://arxiv.org/abs/1507.01760).
  - Gaussian distributions on Riemannian symmetric spaces : statistical learning with structured covariance matrices. Salem Said, Hatem Hajri, Lionel Bombrun, Baba Vemuri, (https ://arxiv.org/abs/1607.06929).

— Version en ligne de l'algorithme EM :
  - Online Expectation-Maximisation. Olivier Cappé, (https ://hal.archives-ouvertes.fr/hal-00532968/document).
  - Online EM algorithm for latent data models. Olivier Cappé, Eric Moulines, (https ://arxiv.org/abs/0712.4273).

— Distance statistique entre clusters :
  - Warped Riemannian metrics for location-scale models. Salem Said, Lionel Bombrun, Yannick Berthoumieu, (https ://arxiv.org/abs/1707.07163).
  - Computing distances and geodesics between manifold-valued curves in the SRV framework.

Alice Le Brigant, (https ://arxiv.org/abs/1601.02358).