Sujet de stage de Master 2 Recherche
Institut de Mathématiques de Bordeaux & INRIA / INSERM - Université de Bordeaux

Mise en correspondance de bio-marqueurs cellulaires par transport optimal de mesures pour l'analyse de données de cytométrie en flux.

Encadrants : Jérémie BIGOT & Boris Hejblum
Institut de Mathématiques de Bordeaux & INRIA / INSERM, Université de Bordeaux
jeremie.bigot@u-bordeaux.fr boris.hejblum@u-bordeaux.fr

Contexte du sujet de stage

La cytométrie en flux est une technologie de mesure à haut débit qui permet de rapidement caractériser un grand nombre de cellules selon leur propriétés physiques et chimiques à partir d'un échantillon biologique. On peut ainsi quantifier des bio-marqueurs cellulaires (intracellulaires ou membranaires) simultanément (c'est-à-dire sous forme de données multivariées) afin de déterminer la sous-population à laquelle appartient chaque cellule. La méthode de référence pour l'analyse de ces données de cytométrie en flux reste aujourd'hui l'approche manuelle, consistant à séparer visuellement les sous-populations cellulaires d'intérêt selon les pics de densité au cours d'un processus appelé *gating* (projections successives du nuage de point en 2 dimensions) et qui s'avère fastidieux, peu reproductible et couteux. Cependant, le développement de cette technologie conduit maintenant à des ensembles de données constitués de mesures multiples (par exemple, jusqu'à 18 bio-marqueurs simultanés) de millions de cellules. Face aux difficultés rencontrées par l'analyse manuelle sur de telles dimensions, un travail important a donc été effectué ces dernières années pour proposer des méthodes statistiques automatiques permettant de surmonter les limitations d'une analyse statistique manuelle [3, 5].

Lors de l'analyse d'échantillons prélevés chez différents patients, un problème crucial dans la cytométrie en flux est la mise en correspondance des populations cellulaires identifiées d'un patient à l'autre. Cette étape est rendue délicate du fait de problèmes d'alignement et de normalisation des données mesurées qui sont liés notamment aux caractéristiques technologiques des appareils d'enregistrement. A titre d'exemple, on peut considérer des données issues de mesures effectuées chez une quinzaine de patients avec une dizaine de bio-marqeurs dans le cadre d'un étude rétrospective de greffe rénale menée par le réseau ITN (Immune Tolerance Network) [4]. Dans la Figure 1, on représente une projection bi-variée de ces mesures à partir des marqueurs cellulaires FSC (forward-scattered light) et SSC (lumière diffusée latéralement) qui permettent de mesurer le volume et la complexité morphologique des cellules. Le nombre de cellules considérées par patient varie de 88 à 2185. Dans la Figure 1, on peut clairement constater un problème d'alignement entre ces mesures qui rend difficile la mise en correspondance de sous-population cellulaires d'un patient à l'autre.

Le principal objectif de ce stage est de mettre en place des méthodes algorithmes et statistiques pour résoudre cette problématique de mise en correspondance de sous-populations de cellules entre différents patients.

Méthodologie statistique

Dans de nombreux problèmes d'apprentissage statistique, il est nécessaire de pouvoir comparer des données organisées sous la forme de nuages de points dans un espace de grande dimension (par exemple l'espace des bio-marqueurs dans la cytométrie en flux).

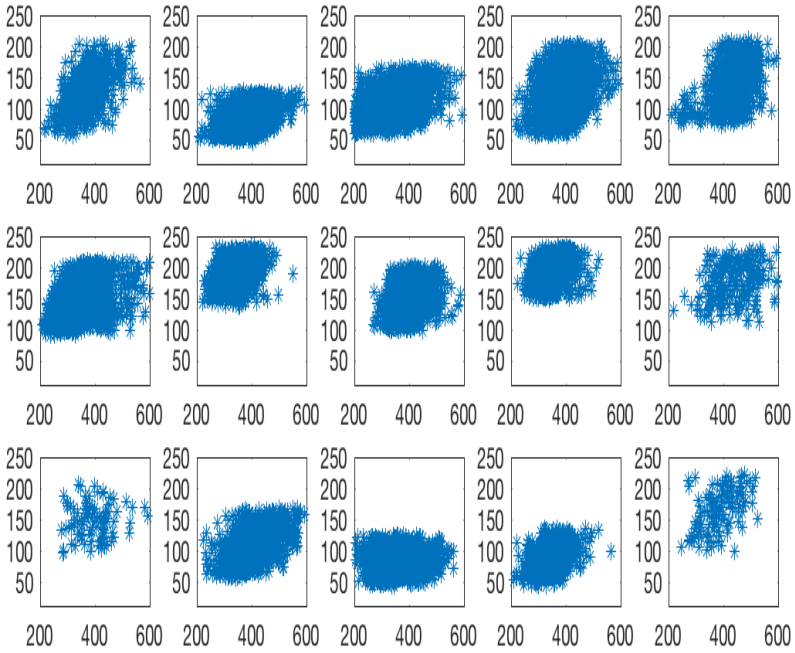

Figure 1. Exemple de données de cytométrie en flux mesurées pour 15 patients (limitée à une projection bi-variée). L'axe horizontal (resp. vertical) représente les valeurs du marqueur cellulaire FSC (resp. SSC).

Il existe bien d'autres exemples d'application qui incluent l'analyse de nuages de mots pour l'étude du langage, la vision par ordinateur, ou bien la catégorisation d'images. L'utilisation de la notion de distance de Wasserstein associée au problème de transport optimal entre des mesures de probabilités est un outil privilégié pour la comparaison de ce type de données qui permet d'atteindre les performances à l'état de l'art pour des applications variées. Pour un aperçu rapide de l'intérêt de l'utilisation de ce type de distance pour l'analyse statistique de données, on pourra consulter cet article de vulgarisation scientifique :

- Bigot, J. (2018) Transport optimal de sable avec une pelle et un seau, nouveaux outils pour l'analyse statistique de données ? *Actualités scientifiques de l'INSMI* :

http://www.cnrs.fr/insmi/spip.php?article2751

Pour une présentation détaillée de nombreux exemples de telles applications et références bibliographiques, on pourra également consulter le livre récent (et tutoriels associés) de Cuturi & Peyré sur les aspects numériques du transport optimal à l'URL :

https://optimaltransport.github.io/

Par ailleurs, un ouvrage de référence très complet sur les aspects mathématiques du transport optimal est le livre [6] de Cédric Villani.

Dans ce stage, il est proposé de s'intéresser à la notion de barycentre dans l'espace de Wasserstein [1, 2] qui permet de généraliser la notion de moyenne usuelle au cas de l'analyse d'un ensemble de nuages de points tels que ceux de la Figure 1. L'un des intérêts

de ce type d'approche statistique est de pouvoir obtenir une mise en correspondance de sous-régions dans un ensemble de nuages de points. Toutefois, en pratique, l'utilisation d'un barycentre de Wasserstein peut être limitée par son coût algorithmique quand la dimension des données augmente. Le principal objectif de ce stage est donc de pouvoir mettre au point des techniques numériques qui rendent faisables ce type d'approche pour l'analyse statistique de données de cytométrie en flux.

Le stage est à l'interface entre la statistique computationnelle et la biostatistique. Les principales notions abordées feront appel à des outils de statistique et d'optimisation avec des applications en bio-informatique. Il nécessite une bonne formation en mathématiques appliquées et statistique (du type Master 2 en Mathématiques Appliquées et Statistique), ainsi que la maîtrise d'un outil de programmation tel que Python.

2. Poursuite en thèse

En fonction de son déroulement, le stage pourra déboucher sur une thèse autour de la thématique "**Analyse de données de cytométrie en flux par transport optimal de mesures**" (sous réserve qu'un financement soit trouvé).

References

[1] Agueh, M., and Carlier, G. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis 43*, 2 (2011), 904–924.

[2] Bigot, J., Cazelles, E., and Papadakis, N. Data-driven regularization of wasserstein barycenters with an application to multivariate density registration. *ArXiv preprint: 1804.08962* (2018).

[3] Commenges, D., Alkhassim, C., Gottardo, R., Hejblum, B. P., and Thiébaut, R. cytometree: A binary tree algorithm for automatic gating in cytometry analysis. *Cytometry Part A* (2018). bioRxiv 335554.

[4] Hahne, F., Khodabakhshi, A., Bashashati, A., Wong, C.-J., Gascoyne, R., Weng, A., Seyfert-Margolis, V., Bourcier, K., Asare, A., Lumley, T., Gentleman, R., and Brinkman, R. Per-channel basis normalization methods for flow cytometry data. *Cytometry Part A 77*, 2 (2010), 121–131.

[5] Hejblum, B. P., Alkhassim, C., Gottardo, R., Caron, F., and Thiébaut, R. Sequential Dirichlet Process Mixtures of Multivariate Skew t-distributions for Model-based Clustering of Flow Cytometry Data. *Annals Of Applied Statistics* (2018). 39 pages, 11 figures.

[6] Villani, C. *Topics in optimal transportation*, vol. 58 of *Graduate Studies in Mathematics*. American Mathematical Society, 2003.