Projet de stage de master 2 : Modèles génératifs des méthodes de réduction de dimension par embedding.

4 novembre 2020

Encadrement : Franck Picard, LBMC ENS Lyon

Contexte applicatif : Des avancées récentes en biologie cellulaire et séquençage à haut débit ont rendu possible l'émergence d'une génomique dite en cellules uniques. Il s'agit désormais d'accéder à l'identité moléculaire d'une population de cellules, sur la base des quantifications individuelles de leur génome, transcriptome, épigénome, et protéome. La variabilité intercellulaire des phénomènes moléculaires est soupçonnée de longue date, mais ce n'est que récemment que l'ampleur des variations peut être mesurée de manière fiable. Les défis méthodologiques posés par la génomique en cellule unique sont majeurs : l'exploitation de ce déluge de données ne pourra être réalisé sans un cadre mathématique et computationnel adapté. D'un point de vue méthodologique, nous avons accès à la distribution de l'expression des gènes sur une population entière, une cellule devenant une composante d'une distribution multidimensionnelle complexe. Aussi, cette variabilité inter-cellulaire mesurée nous informe sur les processus biologiques tels que la régulation des gènes, la différentiation et la prise de décision cellulaire. Le challenge méthodologique est bien devant nous pour appréhender cette complexité jamais rencontrée, et s'accompagne d'un défi computationnel immense pour être en capacité de traiter de tels volumes de données en des temps raisonnables. La réduction de dimension est une étape obligée pour la visualisation des données, mais également pour la simplification des analyses sous jacentes, afin de réduire la complexité des jeux de données qui peuvent atteindre des millions de lignes pour des centaines de milliers de colonnes.

Contexte méthodologique : La réduction de dimension est un sujet central des techniques d'apprentissage statistique, avec les techniques linéaires les plus classiques comme l'analyse en composantes principales. Depuis quelques années de nouvelles méthodes de réduction de dimension non

linéaires ont été développées, comme tSNE (Stochastic Neighbor Embedding) et UMAP (Uniform Manifold Approximation and Projection) afin de proposer de nouvelles représentations qui respectent plus la géométrie locale des nuages de points observés. Ces techniques sont très utilisées et décrites essentiellement par le biais d'algorithmes dont les fondements probabilistes et statistiques sont mal connus. La littérature fait état d'heuristiques largement fondées sur l'utilisation de noyaux (techniques de voisinages), mais la méthode doit être calibrée. Or l'absence de définition satisfaisante d'un cadre statistique des méthodes d'embedding empêche l'exploration de ces aspects de calibration.

Objectif du stage : Lors d'une étude préliminaire, nous avons réussi à mettre au point un modèle génératif pour proposer un cadre statistique satisfaisant pour les méthodes d'embedding comme tSNE et UMAP. Nos analyses sont exploratoires, et il semblerait que le modèle gaussien structuré sur graphe soit une piste de recherche prometteuse. L'objectif du stage est donc de mener à bien cette exploration, pour proposer à terme un cadre statistique complet, ainsi qu'une méthode d'inférence. Les pistes que nous avons déjà commencé à explorer correspondent à des techniques d'inférence probabiliste (modèles hierarchiques, inférence variationnelle), mais la question reste ouverte. Nous nous interrogeons également sur les liens potentiels que notre modèle pourrait avoir avec d'autres méthodes de réduction de dimension non paramétriques comme les variational auto-encoders. Les données étant principalement constituées de comptages, on pourra aussi s'interroger sur la mise au point d'un modèle adapté à ce type de données (discrètes avec hétéroscédasticité). Après avoir défini le modèle statistique, l'objectif du stage sera également de proposer une implémentation efficace de la méthode. Etant donné la complexité des jeux de données, une implémentation rigoureuse et permettant l'execution de la méthode en un temps raisonnable, constitue un véritable objectif. L'ensemble des développements seront effectués sur des données issues des expériences de séquençage à haut débit en cellules uniques. Le stage se déroulera au sein du projet ANR SingleStatomics [1] qui rassemble des spécialistes de machine learning pour les données génomiques.

- UMAP : Uniform Manifold Approximation and Projection for Dimension Reduction, Leland McInnes and John Healy and James Melville, 2020, arxiv 1802.03426
- Stochastic Neighbor Embedding : `https://lvdmaaten.github.io/tsne/`

1. `https://anr-singlestatomics.pages.math.cnrs.fr/`