Sujet de thèse :
ACP fonctionnelle pour données de comptage.

3 décembre 2021

Encadrement de la thèse : .
- Franck Picard (ENS Lyon) - Frank.Picard@ens-lyon.fr
- Vincent Rivoirard (Université Paris Dauphine) - Vincent.Rivoirard@dauphine.fr
- Angelina Roche (Université Paris Dauphine) - Angelina.Roche@dauphine.fr

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. Plus globalement, les données de comptages sont devenues de plus en plus courantes, notamment car beaucoup de systèmes de mesures effectuent des comptages (d'individus, de particules, de séquences). D'un point de vue statistique, l'analyse de comptages en grande dimension pose des problèmes spécifiques, liés par exemple à l'hétéroscédasticité des modèles associés (voir **?**). D'un point de vue pratique, une des difficultés à traiter ce type de données est le manque d'outil de représentation permettant une analyse exploratoire.

Contexte méthodologique : L'objectif de cette thèse est de proposer un cadre statistique général pour analyser et visualiser les données issues des expériences de séquençage à haut débit en cellules uniques. Ces données sont caractérisées par leur nature discrète (ce sont des comptages de quantification), leur forte dispersion, et la présence importante de données manquantes. De plus, nous nous focalisons sur une composante originale des

données de génomique, à savoir leur organisation spatiale le long des chromosomes. Cette organisation crée des dépendances complexes entre les sites voisins, que nous proposons de modéliser dans un cadre fonctionnel adapté. L'objectif est donc de développer un cadre fonctionnel non-paramétrique pour modéliser ces observations, dans un premier temps comme des courbes unidimensionelles le long du génome.

Cadre statistique : Nous nous intéressons à la modélisation statistique de la variabilité au sein d'un échantillon de données de comptage où, pour chaque cellule i, nous observons le signal $(Y_i(t_h))_{h=1,\dots,p}$ des comptages observés aux positions $t_h = h/p$, $h = 1, \dots, p$. Plusieurs modèles existent reliant la loi de $Y_1, \dots, Y_n$ à une suite de fonctions non observées, qui peuvent être aléatoires ou déterministes. C'est le cas par exemple du modèle de Poisson fonctionnel inspiré des travaux de (**?**) où

$$Y_i(t_h) \sim \mathcal{P}(Z_i(t_h)),$$

ou du modèle de Poisson log-normal (**?**)

$$Y_i(t_h) \sim \mathcal{P}(\exp(Z_i(t_h)),$$

où les Z_i sont des variables fonctionnelles i.i.d. Les fonctions $Z_1, \dots, Z_n$ nous donnent des informations précieuses sur la dispersion des données de l'échantillon, l'enjeu est donc de pouvoir les représenter dans un espace de dimension restreinte. Un outil classique pour réduire la dimension d'un échantillon de fonctions aléatoires ou déterministe est l'Analyse en Composantes Principales pour données fonctionnelles (ACPf). L'objectif de la thèse sera de s'appuyer sur les résultats récents en ACPf (**???**) pour développer une procédure d'estimation des fonctions propres de l'opérateur de covariance associé aux courbes $Z_1, \dots, Z_n$. Ces fonctions propres sont aussi appelées composantes principales des données fonctionnelles $Z_1, \dots, Z_n$ et constituent un dictionnaire orthonormé qui nous permet d'obtenir la meilleure représentation des Z_i (décomposition de Karhunen-Loeve).

Le premier axe de recherche consistera à prouver des résultats théoriques sur les estimateurs des composantes principales (inégalités oracles, vitesses de convergence minimax) en s'appuyant sur les travaux existants sur le modèle de Poisson en régression non-paramétrique d'une part (**?**) et sur l'ACPf d'autre part (**???**).

De nombreuses extensions de ce travail seront envisagées, tant d'un point de vue de la modélisation (modèles non Poissoniens, données manquantes), que du point de vue théorique (cas du design fixe ou aléatoire, calibration de pénalités, estimation parcimonieuse).

L'encadrement de la thèse sera précédé d'un stage de Master 2 de 4 mois.

Références