Sujet de thèse : **Statistique nonparamétrique pour des données directionnelles**

Proposé par : Thanh Mai Pham Ngoc (LAGA, Université Sorbonne Paris Nord)
et Claire Lacour (LAMA, Université Gustave Eiffel)

On considère des données circulaires, modélisées par $X_1, \dots, X_n$ des variables aléatoires indépendantes et identiquement distribuées sur le cercle $\mathbb{S}^1$. Ce type de données est fréquemment utilisé en sciences de la terre (direction du vent par exemple) mais aussi en écologie (direction prise par des animaux). Cela permet aussi de modéliser un moment de la journée (ou de la semaine), d'où des applications en médecine. Ces données nécessitent un traitement spécifique ; les méthodes statistiques classiques utilisées pour des variables dans $\mathbb{R}$ ou $\mathbb{R}^d$ ne peuvent plus être utilisées, voir par exemple Mardia and Jupp (2000), ou Pewsey (2004). La littérature sur le sujet se situe majoritairement en statistique paramétrique. Ce sujet de thèse se propose au contraire d'utiliser des outils de statistiques non-paramétriques pour traiter ces questions.

Données censurées sur le cercle ou la sphère

Dans un premier temps, il s'agira de s'intéresser aux cas de données censurées. La censure est une modélisation de données non complètement observées, très utilisée en analyse de survie dans les applications biomédicales. Typiquement, au lieu d'observer directement X_i, on observe seulement un intervalle qui contient X_i. C'est le cas par exemple en épidémiologie ou le moment de contamination ne peut pas être connu de façon exacte. Ce type de censure est appelée censure par intervalle. Un exemple de cas de données censurées sur le cercle avec application en science forensique est donné dans le chapitre 5 de Mulder (2019) : si on s'intéresse à l'heure où se produisent des vols de vélos, on n'a généralement pas accès à l'horaire précis mais seulement à un intervalle pendant lequel il s'est produit. Un premier travail a été réalisé par Jammalamadaka and Mangalam (2009). Le sujet proposé est ici d'établir un estimateur non-paramétrique de la densité dans la lignée des travaux de Brunel (2013). On recherchera l'optimalité de cet estimateur dans le cadre de la théorie minimax. L'adaptation à des régularités inconnues pourra également être étudiée.

Il serait intéressant d'étendre ce travail au cas de données sur la sphère $\mathbb{S}^2$, voire sur l'hypersphère $\mathbb{S}^{d-1}$. Le données sphériques sont particulièrement utilisées en astrophysique, géologie et océanographie. L'étude de données dont on ne connait pas la localisation sphérique précise mais seulement l'appartenance à un petit domaine aurait de nombreuses applications. Dans ce contexte, il s'agira déjà de formuler précisément le problème avant de s'intéresser à l'estimation de la loi. On pourra également étudier comment effectuer des tests statistiques (uniformité, symétrie, adéquation) dans ce contexte de données partielles.

Un modèle de mélange

Un deuxième sujet proposé pour cette thèse est le suivant : on observe des variables i.i.d. $X_1, \dots, X_n$ sur le cercle, ou sur la sphère $\mathbb{S}^{d-1}$, telles que

$$X_i \sim (1-p)f_0 + pf$$

où $p \in]0,1[$ est inconnu, f est une densité inconnue et f_0 est une densité connue. Ce modèle est souvent appelé modèle de contamination. Il apparait naturellement dans le cadre de tests multiples (on s'intéresse à la distribution des p-valeurs, f_0 est la loi uniforme, et p la proportion d'hypothèses alternatives) mais aussi dans des applications directes, comme en astronomie où f_0

peut être vu comme un bruit de fond tandis que f est la distribution du signal d'intérêt, voir par exemple Patra and Sen (2016). Le but est alors d'estimer la densité f ou bien de tester si cette composante inconnue appartient à une famille paramétrique donnée. Le cas de données à valeurs dans $\mathbb{R}$ a été étudié par de nombreux auteurs, voir entre autres Bordes et al. (2006), Nguyen and Matias (2014), Patra and Sen (2016), Pommeret and Vandekerkhove (2019). La transposition d'un modèle de mélange du cas réel au cas circulaire est loin d'être aussi simple qu'il n'y paraît. D'un point de vue mathématique, la topologie du cercle rend le problème très différent du cas linéaire. Ainsi dans le cas du modèle de mélange à 2 composantes inconnues mais translatées l'une de l'autre, Lacour and Pham Ngoc (2022) ont mis en lumière des phénomènes de manque d'identifiabilité propres au cas circulaire. On peut donc se demander si ces phénomènes sont également à l'œuvre dans le cas du modèle de contamination.

Enfin, en supposant que le signal n'est observé que bruité, on pourra s'intéresser au modèle suivant

$$X_i \sim (1-p)f_0 + (1-p)(f_\varepsilon \star f)$$

où f_ε est la densité d'une rotation aléatoire représentant le bruit de mesure, et $\star$ est le produit de convolution. Ce modèle a été étudié dans $\mathbb{R}^d$ par Lepski and Willer (2017) mais en considérant p et f_ε connus.

Références

Bordes, L., Delmas, C., and Vandekerkhove, P. (2006). Semiparametric estimation of a two-component mixture model where one component is known. *Scandinavian journal of statistics*, 33(4) :733–752.

Brunel, E. (2013). *On Adaptive Nonparametric Estimation for Survival Data.* HdR.

Jammalamadaka, S. R. and Mangalam, V. (2009). A general censoring scheme for circular data. *Statistical Methodology*, 6(3) :280–289.

Lacour, C. and Pham Ngoc, T. M. (2022). Semiparametric inference for mixtures of circular data. *Electronic Journal of Statistics*, 16(1) :3482–3522.

Lepski, O. and Willer, T. (2017). Lower bounds in the convolution structure density model. *Bernoulli*, 23(2) :884–926.

Mardia, K. V. and Jupp, P. E. (2000). *Directional statistics.* Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester.

Mulder, K. T. (2019). *Bayesian Circular Statistics von Mises-based solutions for practical problems.* PhD thesis.

Nguyen, V. H. and Matias, C. (2014). On efficient estimators of the proportion of true null hypotheses in a multiple testing setup. *Scandinavian Journal of Statistics*, 41(4) :1167–1194.

Patra, R. K. and Sen, B. (2016). Estimation of a two-component mixture model with applications to multiple testing. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 78(4) :869–893.

Pewsey, A. (2004). Testing for circular reflective symmetry about a known median axis. *Journal of Applied Statistics*, 31 :575–585.

Pommeret, D. and Vandekerkhove, P. (2019). Semiparametric density testing in the contamination model. *Electronic Journal of Statistics*, 13(2) :4743–4793.