Offre de Thèse - projet ModViv : Choix de modèle pour des approches de clustering de données complexes en science du vivant

Contexte : Le doctorant sera supervisé par Marie Du Roy de Chaumaray (Ensai/Crest, Bruz) and Matthieu Marbac (Ensai/Crest, Bruz). La thèse s'effectuera à l'ENSAI. Le doctorant aura pour laboratoire de recherche le CREST. Le financement du projet est acquis. Il s'agit d'un co-financement de la région Bretagne (ARED) et du GENES. Ainsi, la thèse débutera entre le 1er septembre 2023 et le 1er janvier 2024. Le financement est pour une durée de 3 ans.

Candidatures : Cette thèse s'adresse à des étudiants détenant un master ou un diplôme d'ingénieur en Mathématiques appliquées ou en statistiques. Les candidatures (CV, lettre de motivation, 2 lettres de recommandation) doivent être envoyées par email à matthieu.marbac-lourdelle@ensai.fr. Les candidatures seront évaluées au fil du l'eau à partir du **24 juin 2023** jusqu'à ce que le poste soit pourvu.

Résumé synthétique du projet : ModViv est un projet de thèse en statistique dont le but est de proposer des critères de choix de modèles pour la classification non-supervisée. La classification non-supervisée est un outil essentiel de machine learning permettant l'analyse de données massives et complexes, dont le but est de regrouper l'ensemble des individus en quelques classes caractéristiques. ModViv s'intéressera aux approches basées sur des modèles de mélanges dont l'avantage est de se placer dans un cadre probabiliste rigoureux, qui permet le développement d'outils mathématiques pour répondre aux questions primordiales de choix de modèles. Dans ce contexte, les problématiques de choix de modèles ont pour but de déterminer des indicateurs essentiels pour l'analyse (e.g., nombre de groupes de sujets, sous-ensemble de variables discriminantes). ModViv se focalisera sur deux modèles développés pour analyser pour des réseaux d'interactions (Stochastic Block Model) ou des données avec changement de régime (Chaines de Markov Cachées), deux types de données fortement présent en santé et en sciences du vivant. La spécificité de ce projet est d'aborder la problématique du choix de modèle, dans un contexte où la dimension de l'espace des paramètres ainsi que le nombre de modèles concurrents augmentent avec le nombre de sujets à analyser. Cette situation met en défaut les approches actuelles de choix de modèles car celles-ci considèrent que ces deux quantités sont fixes. Pour cela, on s'intéressera aux cas des modèles paramétriques où des résultats de consistance seront établis pour des critères basés sur une pénalisation de la vraisemblance. Dans un second temps, des extensions aux cas des modèles non paramétriques dont le but d'avoir des hypothèses de modélisation les plus réalistes possibles, seront établis principalement en utilisant des discrétisations locales. Enfin, des extensions à l'apprentissage online, qui est un challenge majeur du deep-learning, seront proposées.

Projet : Pour analyser des données massives, les modèles de mélange permettent d'atteindre l'objectif de classification non-supervisée en définissant un cluster comme l'ensemble des observations issues de la même composante du mélange. En se plaçant dans un cadre probabiliste, ces modèles permettent l'utilisation de critères basés sur la vraisemblance pénalisée, pour répondre à des questions cruciales de choix de modèles (e.g., estimation du nombre de groupes). L'utilisation de ces critères nécessite le contrôle du supremum du rapport de vraisemblance (LRT) qui peut être obtenu de manière asymptotique par une reparamétrisation "locally conic" [1, 2]. Ce contrôle n'étant uniquement asymptotique, ces outils nécessitent que l'espace des paramètres ainsi que le nombre de modèles en concurrence doivent être fixe [3]. De plus, le contrôle du LRT reste une question ouverte pour des données dépendantes. Enfin, jusqu'au travaux de [4], les approches basées sur la vraisemblance pénalisée nécessitaient des hypothèses de distributions paramétriques. Cependant ces hypothèses peuvent entraîner des biais importants lorsqu'elles sont violées. Ainsi, [4] proposent des outils de choix de modèles pour des mélanges non paramétriques en combinant des discrétisations locales permettant de revenir à un modèle paramétrique particulier et un contrôle non-asymptotique du LRT. L'idée centrale de ModViv est d'utiliser le contrôle non-asymptotique du LRT obtenu dans [4] pour deux modèles cruciaux en science du vivant. Ainsi, ce projet se découpe en une étape préliminaire et deux étapes principales. Dans un premier temps, un critère de choix de

modèle sera proposé pour le cas où le nombre de modèles et où l'espace des paramètres augmentent avec la taille de l'échantillon. La consistance de ce critère sera basée sur le contrôle non-asymptotique du LRT de [4]. Dans un second temps, on s'intéressera à l'estimation du nombre de groupes dans un réseau d'interactions modélisé par un Stochastic Block Model (SBM). Ces modèles sont utiles pour modéliser l'interaction entre gènes ou entre espèces. Dans ce contexte, on cherche à estimer des groupes de gènes (ou d'espèces) qui interagissent plus fortement entre eux. Pour ce modèle, aucun critère de choix de modèle consistant n'a pu être établi. L'absence de critère de choix de modèle pour le SBM s'explique par la forme complexe du LRT qui empêche l'utilisation des techniques standard permettant son contrôle asymptotique. De plus, pour ces modèles, il est d'usage de considérer que le nombre de groupes peut augmenter avec la taille d'échantillon. En utilisant la convergence de la fonction de vraisemblance complétée vers la fonction de vraisemblance combinée avec le contrôle non asymptotique du LRT, nous pourrons établir la consistance d'un critère basé sur la vraisemblance complétée pénalisée. De plus, cette approche pourra autoriser le nombre maximum de groupes à croître avec la taille d'échantillon. Dans un troisième temps, on s'intéressera à l'estimation de l'ordre d'une chaîne de Markov cachée (HMM) avec des lois d'émissions non-paramétriques. Les HMM sont une généralisation des modèles de mélange qui relâchent l'hypothèse d'indépendance entre les observations. Ce sont des outils puissants de modélisation, notamment en écologie, car ils permettent de considérer des changements de régimes. En utilisant des outils de "coupling" pour gérer la dépendance entre observations, le contrôle non asymptotique du LRT d'un HMM pourra être obtenu, dans le cas de lois d'émissions paramétriques. Ensuite, en utilisant des discrétisations locales similaires à [4], une approche simple de l'estimation de l'ordre d'un HMM paramétrique pourra être obtenu. De plus, cette approche pourra autoriser une augmentation de l'espace des modèles ce qui permettra une estimation "online" de l'ordre d'un HMM.

Environnement scientifique Le doctorant dépendra de l'école doctorale MATISSE pour obtenir le titre de Docteur en Mathématiques Appliqués. Il sera affecté à l'École Nationale de la Statistique et de l'Analyse de l'Information (ENSAI) située sur le campus de Ker Lann à Bruz (35172 - France). L'ENSAI est une grande école formant des ingénieurs en data science dont la compétence majeure est la modélisation statistique à laquelle sont associées des compétences en informatique et en économie quantitative. L'ENSAI à une activité forte dans la bassin régional puisqu'elle est membre de deux EUR (DIGISPORT et CYBER). Plus précisément, le doctorant sera affecté à l'équipe statistique de l'ENSAI qui est composée de 17 enseignants chercheurs et de 9 doctorants. La taille de l'équipe lui assure un environnement où le doctorant pourra s'épanouir. Celui-ci se verra offrir la possibilité de dispenser des enseignements en statistique à l'ENSAI, ce qui lui permettra d'obtenir une expérience solide d'enseignement qui sera reconnue lors d'une future candidature à un poste académique. De plus, les collaborations existantes des membres de l'équipe avec des équipes pluridisciplinaires (AgroCampusOuest, INSERM Paris et Faculté de Médecine de Lille) assureront que les développements statistiques effectués lors de la thèse seront fait dans le but d'être ensuite applicables sur des analyses de données réelles, notamment issues du monde du vivant. Le doctorant aura pour laboratoire le Centre de recherche en économie et en statistique (CREST). Le CREST est un laboratoire commun à l'ENSAE Paris, à l'ENSAI et au Département d'Économie de l'École polytechnique, qui fédère et dynamise une activité de recherche prolifique tant fondamentale qu'appliquée dans les domaines de la statistique, de l'économie et de l'informatique. Le CREST est composé de plusieurs statisticiens renommés, ce qui lui assure une visibilité forte au niveau international. Cette visibilité sera sans nulle doute bénéfique pour la diffusion des travaux du doctorant.

Références

[1] Dacunha-Castelle, D., & Gassiat, E. (1997). *Testing in locally conic models, and application to mixture models.* ESAIM : Probability and Statistics, 1, 285-317.

[2] Dacunha-Castelle, D., & Gassiat, E. (1999). *Testing the order of a model using locally conic parametrization : population mixtures and stationary ARMA processes.* The Annals of Statistics, 27(4), 1178-1209.

[3] Keribin, C. (2000). *Consistent estimation of the order of mixture models.* Sankhyā : The Indian Journal of Statistics, Series A, 49-66.

[4] Du Roy de Chaumaray, M & Marbac, M. (2022). *Full Model Estimation for Non-Parametric Multivariate Finite Mixture Models.* arXiv, 2112.05684.