Sujet de stage de M2 en statistique

Présentation

Sujet.
Clustering dans un modèle de graphes aléatoires à positions latentes

Encadrants.
Gaëlle Chagny (gaelle.chagny@univ-rouen.fr) et Antoine Channarond (antoine.channarond@univ-rouen.fr), Laboratoire de Mathématiques Raphaël Salem, UMR CNRS 6085, Université de Rouen Normandie.

Financement et contexte. ANR SMILES (Statistical Modeling and Inference for unsupervised Learning at largE-Scale, ANR-18-CE40-0014).

Ce projet est principalement dévolu à la modélisation et l'inférence statistique pour des données complexes, de grande échelle ("Big data"), via des problèmes de régression et de classification non-supervisée. Le sujet du stage s'insère dans la partie graphe aléatoire du projet.

Durée. 4 mois (à partir de mars ou avril 2022).

Résumé

Mots Clés. Graphes aléatoires. Modèles à variables latentes. Clustering.

Contexte.

Le sujet de ce stage se situe dans le domaine de l'analyse statistique des graphes aléatoires. L'essor de ce domaine de la statistique remonte à la seconde moitié du XXème siècle, suite aux travaux fondateurs d' Erdős et Rényi (1959), Erdős et Rényi (1960), et encore Erdős et Rényi (1961). La modélisation mathématique d'un réseau d'interaction entre individus, au sens d'un ensemble d'interactions entre les membres d'un groupe, se fait généralement par un graphe : les membres du réseau sont les noeuds, et les interactions sont représentées par les paires de noeuds, les arêtes. L'inférence statistique des graphes aléatoires trouve donc des applications dans de multiples domaines : sans être exhaustif, on peut citer la sociologie (étude des liens sociaux établis entre individus d'un groupe, étude de la propagation d'une rumeur dans un réseau social, ...), la biologie (étude de la propagation d'une épidémie dans une population, étude d'un système de régulation de protéines ou de gènes...), ou encore l'informatique (réseaux de partage de données, réseaux de page web reliées par des liens hypertextes...).

Objectifs du stage.

Motivé par ces nombreuses applications, l'inférence statistique des graphes est aujourd'hui l'objet de nombreuses recherches. On s'intéressera à la question de la recherche d'une structure (ou clustering) dans les noeuds d'un graphe aléatoire à variables latentes c'est à dire non observées. Dans cette catégorie de modèles, on attribue à chaque noeud i une variable aléatoire non observée (latente, donc) Z_i, et la probabilité de connexion des noeuds i et j dépend conditionnellement de Z_i et Z_j . La structure d'un tel modèle statistique est décrite par les variables latentes et leur loi, et l'inférence se fait à partir du graphe, seule variable observée.

Le modèle que l'on propose d'étudier dans ce stage, principalement en travaillant sur le chapitre 7 de la thèse de Channarond (2013), est le *Latent Position Cluster*. Dans ce modèle, introduit par Handcock *et al.* (2007), les variables latentes Z_i sont des positions dans l'espace $\mathbb{R}^d$, admettant une

densité f par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$, et la probabilité de connexion de deux noeuds i et j dépend de la distance entre leurs positions. Les clusters sont définis comme les composantes connexes de l'ensemble de niveau $\{f \geq t\}$ fixé de f, et l'objectif est d'en estimer le nombre à partir du graphe. Il est possible d'estimer la densité en les positions latentes des noeuds grâce à leur degré, ce qui permet d'établir une correspondance entre les clusters et les composantes connexes de certains sous-graphes du graphe observé, obtenus en retirant les noeuds de faible degré. En particulier, un estimateur du nombre de clusters peut en être déduit, et sa consistance en un certain sens démontrée. On obtient ainsi un algorithme d'inférence. Enfin, il sera possible d'étudier les problèmes de sous- et sur-estimation du nombres de clusters : si la non sous-estimation est un problème résolu, grâce à des outils d'estimation non-paramétrique de densité, la sur-estimation du nombre de composantes reste un problème au moins partiellement ouvert, qui pourra faire l'objet d'un travail de recherche en fin de stage.

La bibliographie ci-dessous donne quelques éléments indicatifs.

Références

G. Biau, B. Cadre et B. Pelletier : A graph-based estimator of the number of clusters. *ESAIM : Probability and Statistics*, 11:272–280, 2007.

A. Channarond : *Recherche de structure dans un graphe aléatoire : modèles à espace latent.* Thèse de doctorat, Université Paris Sud-Paris XI, 2013.

F. Comte : *Estimation non-paramétrique.* Spartacus IDH, 2015.

P. Erdős et A. Rényi : On random graphs. I. *Publ. Math. Debrecen*, 6:290–297, 1959. ISSN 0033-3883.

P. Erdős et A. Rényi : On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 5:17–61, 1960. ISSN 0541-9514.

P. Erdős et A. Rényi : On the strength of connectedness of a random graph. *Acta Mathematica Hungarica*, 12(1):261–267, 1961.

M. S. Handcock, A. E. Raftery et J. M. Tantrum : Model-based clustering for social networks. *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, 170(2):301–354, 2007.

A. B. Tsybakov : *Introduction to nonparametric estimation.* Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.