Offre de thèse en Mathématiques Appliquées - Statistique:

Modélisation et inférence statistiques de réseaux dynamiques à l'échelle

Statistical modelling and inference of large-scale dynamic networks

Sujet:

Les réseaux permettent de représenter des structures de données d'interactions entre les éléments d'un système complexe. Ils sont utilisés dans plusieurs domaines, en allant de la biologie (e.g., réseaux de neurones, réseaux de régulation génétique, réseaux métaboliques), jusqu'aux sciences humaines et sociales, notamment les réseaux sociaux (e.g., Twitter, LinkedIn, Facebook).

L'analyse statistique de données de type réseau est devenue un élément essentiel pour inférer l'organisation du réseau et comprendre les interactions entre ses éléments et ce à partir de données brutes. Les modèles statistiques les plus attrayants considèrent le réseau comme un graphe aléatoire, parmi lesquels on peut citer les modèles stochastiques à blocs latents (SBM) (Snijders and Nowicki, 1997; Nowicki and Snijders, 2001; Daudin et al., 2008; Yang et al., 2011; Latouche et al., 2011; Celisse et al., 2012; Ambroise and Matias, 2012; Matias and Robin, 2014; Zreik et al., 2016; Bouveyron et al., 2016; Matias and Miele, 2016)). D'autres modèles à variables latentes se basent sur les mélanges d'experts (Gormley and Murphy, 2010, 2011).

Ces modèles statistiques de graphes aléatoires sont utilisés dans différents domaines d'applications, comme pour les réseaux biologiques (Picard et al., 2009), en sciences humaines et sociales (Gormley and Murphy, 2010, 2011), comme l'analyse de réseaux sociaux (Yang et al., 2011), l'analyse d'une scène politique (Latouche et al., 2011), l'analyse de réseaux de données textuelles (Xu and Hero, 2013) notamment pour la classification de topics (Bouveyron et al., 2016), l'analyse de réseaux de données maritimes (Zreik et al., 2016), etc.

La majorité des travaux sur le sujet concernent une modélisation statique du réseau, avec un objectif principalement de clustering en se basant sur un modèle stochastique à blocs latents statique. Or il est clair que la prise en compte de la dynamique des interactions au sein d'un réseau est une question majeure pour pouvoir restaurer et comprendre des propriétés évolutives des interactions entre les éléments du réseau. Dans ce cadre, on peut citer des travaux récents proposant une modélisation dynamique (Yang et al., 2011; Xu and Hero, 2013; Zreik et al., 2016; Matias and Miele, 2016).

Le travail attendu de cette thèse est d'étudier les modèles statistiques pour les réseaux et de proposer de nouveaux modèles statistiques dynamiques (à blocs latents, de mélanges d'experts, de réseaux profonds (hiérarchiques)) pour la modélisation de réseaux évolutifs sous forme de graphes aléatoires dynamiques. Avec la volumétrie des données actuelles, l'objectif est également de proposer des algorithmes d'inférence efficaces dans un contexte à large-échelle (Big Data). L'accent sera mis en particulier sur l'apport d'une rigueur et d'un formalisme statistique aux modèles et aux algorithmes développés. L'aspect appliqué concernerait notamment l'analyse de réseaux sociaux impliquant notamment des données textuelles évolutives (tweets, etc) pour le suivi de la dynamique de communautés et de réseaux biologiques impliquant notamment des données fonctionnelles temporelles.
Parmi les travaux de l'équipe en lien avec le sujet et portant notamment sur des modèles dynamiques à variables latentes et d'analyse de données temporelles, on peut citer Chamroukhi (2016b,a, 2015); Chamroukhi et al. (2013a,b); Samé et al. (2011); Chamroukhi et al. (2009).

Mots clés: réseaux, clustering de graphes, graphes aléatoires dynamiques, modèles à blocs latents, méthodes variationnelles, algorithmes EM, modèles de Markov cachés, inférence à l'échelle, calcul parallèle.

Encadrement:

Directeur de thèse: Faïcel Chamroukhi, Professeur. http://math.unicaen.fr/∼chamroukhi/
Laboratoire: Laboratoire de Mathématiques Nicolas Oresme (LMNO) - UMR CNRS 6139
Etablissement: Université de Caen-Normandie, Caen, France.
Financement: Bourse ministérielle (∼1400 euros nets/mois)

Environnement de travail:

Le **LMNO** est une Unité Mixte de Recherche (UMR) de l'université de Caen et du CNRS. Le LMNO est un laboratoire de très haut niveau, avec une très forte visibilité et reconnaissance internationales. Tous les indicateurs scientifiques sont très hauts. Il regroupe l'ensemble des chercheurs en mathématiques et applications à l'université de Caen et dans les établissements rattachés, soit une cinquantaine de permanents. Les thématiques du LMNO couvrent un spectre large allant du plus fondamental au plus appliqué et sont déclinées au sein de cinq équipes dont l'équipe APS (analyse, probabilités et statistiques) au sein de laquelle s'effectuera la thèse.
L'**Université de Caen Normandie (UNICAEN)** accueille plus de 28 000 étudiants et emploi près de 1580 enseignants-chercheurs et enseignants sur plusieurs campus dans l'agglomération caennaise et en région. Fondée en 1432, par le roi d'Angleterre Henri VI, l'UNICAEN est l'une des plus anciennes universités françaises. Son campus 1 (à Caen) est classé aux monuments historiques. Parmi les grands savant célèbres qui furent à l'UNICAEN, on peut citer Henri *Poincaré* qui y fut enseignant et Pierre-Simon *Laplace* qui y fut étudiant. UNICAEN profite des atouts de la région comme le littoral, les échanges transmanche, une voie rapide vers la capitale et propose des actions culturelles et des activités sportives attrayantes. UNICAEN bénéficie également d'une tradition culturelle forte qui s'est d'ailleurs illustrée en 2014 par de grandes manifestations internationales, le 70e anniversaire du Débarquement et les Jeux équestres mondiaux.

Profil recherché:

Diplôme requis: Etre titulaire d'un Master recherche en mathématiques appliquées (statistique) ou disciplines proches (e.g., Apprentissage automatique, Traitement statistique du signal), depuis moins de deux ans à la date de Juin 2017.
Compétences requises: *i)* des compétences théoriques en modélisation et inférence statistique; *ii)* maîtrise de l'un des langages de programmation suivants: Matlab, R, Python; *iii)* un très bon niveau en anglais.
Compétences souhaitées: Apprentissage statistique non-supervisé, méthodes variationnelles, inférence Bayésienne, utilisation de plateformes BigData (MapReduce, Hadoop, Spark, cloud computing).

Mode de candidature:

Dossier à envoyer sous la forme d'**UN SEUL DOCUMENT .pdf** contenant les pièces suivantes : CV + lettre de motivation + relevés de notes des trois dernières années + tout autre éventuel document (lettres de recommandation, publications scientifiques, ...) à chamroukhi@unicaen.fr.

References

Ambroise, C. and Matias, C. (2012). New consistent and asymptotically normal parameter estimates for random-graph mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):3–35.

Bouveyron, C., Latouche, P., and Zreik, R. (2016). The stochastic topic block model for the clustering of vertices in networks with textual edges. *Statistics and Computing*, pages 1–21.

Celisse, A., Daudin, J.-J., and Pierre, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, 6:1847–1899.

Chamroukhi, F. (2015). Unsupervised learning of regression mixture models with unknown number of components. *Journal of Statistical Computation and Simulation*. Published online: 05 Nov 2015.

Chamroukhi, F. (2016a). Piecewise regression mixture for simultaneous functional data clustering and optimal segmentation. *Journal of Classification*, 33(3):374–411.

Chamroukhi, F. (2016b). Robust mixture of experts modeling using the t distribution. *Neural Networks*, 79:20 – 36.

Chamroukhi, F., Glotin, H., and Samé, A. (2013a). Model-based functional mixture discriminant analysis with hidden process regression for curve classification. *Neurocomputing*, 112:153–163.

Chamroukhi, F., Samé, A., Govaert, G., and Aknin, P. (2009). Time series modeling by a regression approach based on a latent process. *Neural Networks*, 22(5-6):593–602.

Chamroukhi, F., Trabelsi, D., Mohammed, S., Oukhellou, L., and Amirat, Y. (2013b). Joint segmentation of multivariate time series with hidden process regression for human activity recognition. *Neurocomputing*, 120:633–644.

Daudin, J.-J., Picard, F., and Robin, S. (2008). A mixture model for random graphs. *Statistics and Computing*, 18(2):173–183.

Gormley, I. C. and Murphy, T. B. (2010). A mixture of experts latent position cluster model for social network data. *Statistical methodology*, 7(3):385–405.

Gormley, I. C. and Murphy, T. B. (2011). Mixture of experts modelling with social science applications. *Journal of Computational and Graphical Statistics*, 19(2):332–353.

Latouche, P., Birmelé, E., and Ambroise, C. (2011). Overlapping stochastic block models with application to the french political blogosphere. *The Annals of Applied Statistics*, 5(1):309–336.

Matias, C. and Miele, V. (2016). Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages n/a–n/a.

Matias, C. and Robin, S. (2014). Modeling heterogeneity in random graphs through latent space models: a selective review. *ESAIM: Proc.*, 47:55–74.

Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087.

Picard, F., Miele, V., Daudin, J.-J., Cottret, L., and Robin, S. (2009). Deciphering the connectivity structure of biological networks using mixnet. *BMC Bioinformatics*, 10(6):S17.

Samé, A., Chamroukhi, F., Govaert, G., and Aknin, P. (2011). Model-based clustering and segmentation of time series with changes in regime. *Advances in Data Analysis and Classification*, 5(4):1–21.

Snijders, T. A. and Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1):75–100.

Xu, K. S. and Hero, A. O. (2013). *Dynamic Stochastic Blockmodels: Statistical Models for Time-Evolving Networks*, pages 201–210. Springer Berlin Heidelberg, Berlin, Heidelberg.

Yang, T., Chi, Y., Zhu, S., Gong, Y., and Jin, R. (2011). Detecting communities and their evolutions in dynamic social networks—a bayesian approach. *Machine Learning*, 82(2):157–189.

Zreik, R., Latouche, P., and Bouveyron, C. (2016). The dynamic random subgraph model for the clustering of evolving networks. *Computational Statistics*, pages 1–33.