Estimation dans les modèles de graphes à espace latent continu de type MMBM

Jean-Benoist Leger [1]

[1] *MaIAGE, UR 1404 INRA, Jouy-en-Josas, France,* jbleger@jouy.inra.fr

Résumé. La représentation sous forme de réseaux permet de présenter des données divers et variées entre des éléments. Ces données peuvent être binaires, comme une présence/absence de relations, quantifiées, continues, où être valuées sur d'autres espaces. Il est nécessaire de formuler des hypothèses parfois contraignantes pour proposer un modèle probabiliste adapté ces données. Des hypothèses d'appartenance des nœuds à des classes latentes et une indépendance de la loi sur les liens conditionnellement à la loi sur les nœuds conduit à des modèles comme Stochastic Block Model (SBM). Il est possible de relacher la contrainte sur l'espace latent comme dans l'Overlapping Stochastic Block Model ou le Mixed Membership Block Model qui offre plus de libertés avec un espace latent continu. Il est également possible de disposer d'information extérieures pouvant être introduites sous forme de covariables.

Cette présentation introduira les modèles de graphes à classes latentes Mixed Membership Block Model avec ou sans covariables pour diverses lois de probabilités sur les liens. Elle présentera l'extension à la classe latente continue, et introduira une méthode d'estimation basée sur le Variational-EM.

Mots-clés. Graphes aléatoires, Clustering, Stochastic Block Model, Mixed Membership Block Model

Abstract. Various data can be written as networks. These data can be binary, such as a presence/absence information, quantified, continuous or valued in arbitrary space. To build probabilistic model which modelling this data, hypothesis are necessary. With hypothesis as node membership into latent classes, and connectivity behavior which depends only on classes, we obtain models like Stochastic Block Model (SBM). To drop constraints on latent space which is a discrete one where a node only behave on one group, as done for Overlapping SBM, or Mixed Membership BM (MMBM) which have a continuous latent space. Also, it is possible to take care of external information using covariates.

This talk will introduce graph models with latent class of Mixed Membership Block Model type, with or without covariates, with arbitrary distribution on edges, and it will introduce a inference method based on a Variational-EM

Keywords. Random graphs, Clustering, Stochastic Block Model, Mixed Membership Block Model

1 Introduction

Lorsqu'il existe des données entre couples d'éléments, qu'elles soient binaires, quantifiées ou valuées sur d'autres espaces, la représentation d'icelles sous forme de réseaux est un outil d'analyse efficace.

Cette vue de l'esprit est particulièrement adaptée pour représenter des données biologiques, comme dans les cas d'interactions entre protéines ou gènes, des données écologiques, avec des interactions entre des individus, ou des données sociologiques comme dans le cas des réseaux d'interactions sociales.

Modéliser et analyser la structure des réseaux permet de déduire des informations de haut niveau et de réduire la complexité de l'analyse d'un sous-ensemble de caractéristiques. Un mode d'analyse classique est constitué par le clustering des nœuds d'un graphe, et un modèle pour classer les nœuds d'un graphe peut être le Stochastic Block Model (SBM). Introduit par Nowicki and Snijders [2001] il modélise les liens conditionnellement à l'appartenance des nœuds à une classe non observée qui sera prédite par le processus d'inférence. Ce modèle peut s'étendre à des valeurs arbitraires sur les liens [Mariadassou et al., 2010].

Toutefois pour certaines données, comme celles étudiées par Lagache et al. [2013], il apparait nécessaire de vouloir relacher les contraintes d'appartenance stricte à une classe et d'autoriser l'appartenance comme une notion floue entre plusieurs classes. Le modèle Overlapping SBM introduit par Latouche et al. [2011] permet d'appartenir à plusieurs classes, mais présente le défaut de ne pas autoriser d'appartenir à n'importe quel degré de mélange entre les classes. Pour appartenir à n'importe quel degré d'appartenance entre les classes, Daudin et al. [2010] propose d'introduire un ensemble de paramètres sur un simplex entre les classes. Cette dernière approche, même si permet de résoudre le problème d'appartenance, pose des problèmes d'identifiabilité et n'est pas généralisable pour des lois arbitraires et dans le cas de covariables.

Airoldi et al. [2009] proposent quant à eux d'introduire un modèle, qu'ils nomment Mixed Membership Block Model (MMBM) avec une première couche latente sur le même simplex que les paramètres de Daudin et al. [2010], tout en conservant une seconde couche latente avant la couche observé.

Je propose une extensions des modèles MMBM avec diverses lois de probabilités sur les liens, avec des covariables optionnelles., Je propose également une méthode d'estimation basées sur un EM variationnel adaptée à ces modèles.

2 Présentation du modèle

Le modèle doit permettre d'avoir une description continue du comportement de chaque nœud. Pour atteindre ce but, chaque nœud est décrit par une variable latente continue correspondante à la proportion de chaque comportement qu'elle adopte. Cette première couche sera notée $\boldsymbol{W}$.

Pour chaque lien possible, chaque nœud va adopter un comportement de liaison. Cette réalisation ne va dépendre que du mélange de comportement caractéristique de chaque nœud généré à la couche précédente.

Ensuite connaissant les comportements adoptés par chacun des nœuds pour chaque lien, la loi sur le lien ne dépendra alors que des comportements adoptés par les deux nœuds pour se connecter.

Introduisons quelques notations, soit :

— n, le nombre de nœuds ;
— Q, le nombre de comportements ;
— $\boldsymbol{W}$ les mélanges de comportements caractéristiques de chaque nœuds ;
— $\boldsymbol{Z}$ les comportements adoptés par les nœuds pour se connecter entre nœuds ;
— $\boldsymbol{X}$ la matrice d'adjacence.

Le modèle présenté sera donc le suivant :

$$\begin{cases} \boldsymbol{W}_i \sim \mathcal{D}(\boldsymbol{\alpha}), & i \in [\![1, n]\!] \\ \boldsymbol{Z}_{ij} | \boldsymbol{W}_i \sim \mathcal{M}(1, W_i), & i, j \in [\![1, n]\!] \\ X_{ij} | Z_{ijq} Z_{jil} = 1 \sim \mathcal{F}_{ql}, & i, j \in [\![1, n]\!],\ q, l \in [\![1, Q]\!] \end{cases}$$

$\boldsymbol{\alpha}$ étant le paramètre propre à la première couche latente, et $\mathcal{F}_{ql}$ étant une loi connue (mais paramétrée, de paramètres à estimer).

3 Inférence

Pour réaliser l'inférence sous ce modèle, c'est à dire pour estimer les paramètres et pour prédire les couches latentes, une stratégie de la maximisation de la vraisemblance sera présenté.

3.1 EM et approximation variationnelle

La vraisemblance incomplète n'étant pas calculable en un temps raisonnable, comme dans nombre de modèles à couches latentes, une stratégie basée sur l'Expectation–Maximization algorithm (EM) [Dempster et al., 1977] sera présentée.

L'étape E n'étant pas calculable non plus en un temps raisonnable, une stratégie variationnelle [Jaakkola, 2001] est utilisée et sera présentée.

L'EM avec approximation variationnelle se résumant donc à une maximisation alternée avec les paramètres variationnels d'un coté et les paramètres originaux de l'autre.

Cette approximation variationnelle introduit de nombreux paramètres qui feront l'objet d'une explication détaillée dans la mise en œuvre technique.

3.2 Mise en œuvre technique

L'implémentation de cette méthode d'inférence doit se faire en considérant la dimension du problème.

La maximisation par rapport aux paramètres originaux ne pose pas de problèmes particuliers, elle peut se faire explicitement ou au moyen d'un algorithme de quasi-Newton selon la loi utilisée, ces cas seront détaillés lors de la présentation.

La maximisation par rapport aux paramètres variationnels n'est pas faisable par une méthode d'itération d'une équation du point fixe, contrairement aux modèles de type SBM. Une méthode de quasi-Newton stockant une approximation du Hessien n'est pas utilisable au vu de la dimension du problème.

Le choix adopté a été d'effectuer une reparamétrisation suivi d'un algorithme de quasi-Newton stockant une version approché du Hessien sous forme implicite (L-BFGS), ce choix sera présenté et détaillé dans la présentation.

L'initialisation joue un rôle important dans les méthodes de type EM, toutefois, cette méthode est relativement robuste au choix de l'initialisation. L'absolute eigenvalues spectral clustering a été choisi, cette méthode présentant de bonnes propriétés [Rohe et al., 2011] dans le cas du modèle SBM dont MMBM est le prolongement. Ceci sera illustré pendant la présentation.

4 Application

Cette méthode a été utilisée sur des données de réseau, en simulation et sur données réelles.

4.1 En simulation

Dans le premier cas des données ont été générées suivant le modèle et la méthode d'inférence a été appliquée. Ces résultats valident la méthode d'inférence et l'importance mineure de la méthode d'initialisation.

De plus des données ont été simulés sous un autre modèle, la méthode a été utilisée. Nous constatons que les résultats restent cohérents avec l'attendu.

4.2 Données réelles

Cette méthode a été utilisée sur un jeu de données biologique déjà analysé. Les résultats seront présentés et commentés.

Références

E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. In *Advances in Neural Information Processing Systems*, pages 33–40, 2009.

J.-J. Daudin, L. Pierre, and C. Vacher. Model for heterogeneous random networks using continuous latent variables and an application to a tree–fungus network. *Biometrics*, 66(4) :1043–1051, 2010.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

T. S. Jaakkola. 10 Tutorial on Variational Approximation Methods. *Advanced mean field methods : theory and practice*, page 129, 2001.

L. Lagache, J.-B. Leger, J.-J. Daudin, R. J. Petit, and C. Vacher. Putting the biological species concept to the test : Using mating networks to delimit species. *PloS one*, 8(6) : e68267, 2013.

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

M. Mariadassou, S. Robin, and C. Vacher. Uncovering latent structure in valued graphs : a variational approach. *The Annals of Applied Statistics*, pages 715–742, 2010.

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

K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, pages 1878–1915, 2011.