Consistance des estimateurs variationnels pour un modèle de graphe aléatoire

Alain Celisse, Jean-Jacques Daudin

CONSISTANCE DES ESTIMATEURS VARIATIONNELS POUR UN MODÈLE DE GRAPHE ALÉATOIRE

Alain Celisse & Jean-Jacques Daudin

UMR 8524 CNRS – Université Lille 1
Laboratoire de Mathématiques Paul Painlevé
59655 Villeneuve d'Ascq Cedex - France
alain.celisse@math.univ-lille1.fr

UMR AgroParisTech/INRA518
AgroParisTech
75231 Cedex 05, Paris - France
SSBgroup
jean-jacques.daudin@agroparistech.fr

Résumé
Les modèles statistiques pour les graphes aléatoires hétérogènes sont utilisés dans beaucoup de domaines : réseaux sociaux, réseaux écologiques, réseaux biologiques. Parmi ces modèles, les modèles de mélange sur les noeuds servent à identifier les sous-groupes de noeuds ayant une connectivité similaire, ce qui permet d'analyser la topologie du graphe. Dans ce cadre, la vraisemblance n'est cependant pas calculable et l'algorithme EM doit être remplacé par une version variationnelle dont les propriétés n'étaient pas établies jusqu'à présent. Dans cet exposé nous démontrons la consistance des estimateurs variationnels.

Abstract
Statistical models for random graphs are involved in numerous areas such as social, ecological, and biological networks. Thus, mixture models are a widespread tool for identifying subgroups in the graph with nearly the same connectivity. For instance, this enables the analysis of the underlying topology of the graph. However, evaluating the resulting likelihood is computationally intractable. The EM algorithm has to be modified on the basis of a *variational approximation*. The goal of the present talk is to derive the statistical properties of this procedure. In particular, the consistancy of the variational estimators is established.

Mots-clés : Statistique mathématique, Modèles semi et non paramétriques

1 Introduction

Les réseaux complexes sont de plus en plus étudiés en sciences sociales, écologie et biologie moléculaire. On cherche souvent une représentation simplifiée basée sur une classification non supervisée des noeuds. Ceci permet d'obtenir un réseau réduit comportant quelques

méta-noeuds, reliés par des *meta-arcs* caractérisant la topologie du graphe (voir [4]). Par exemple le "clustering" de graphe cherche à obtenir des sous-groupes avec beaucoup de connections intra-groupes et peu de connections inter-groupes (voir [8]).

Le modèle de graphe aléatoires Mixnet ([2],[3]) est bien adapté à cet objectif. Cependant l'estimation des paramètres pose des problèmes difficiles : les méthodes MCMC demandent un calcul intensif et ne permettent pas d'analyser plus de 200 noeuds [3]. De plus, les propriétés de l'*approximation variationnelle* utilisée par [2], permettant d'analyser des graphes jusqu'à 10000 noeuds (voir package Mixnet *http://stat.genopole.cnrs.fr/software/mixnet/*), sont encore mal connues. En particulier, la consistance des estimateurs variationnels reste un problème ouvert dans de nombreuses situations.

Dans cet exposé, nous étudions les propriétés asymptotiques des estimateurs variationnels des paramètres pour le modèle Mixnet, ainsi que pour un modèle semi-paramétrique lié à Mixnet.

2 Mixtnet : modèle de mélange pour les graphes aléatoires

On considère un graphe comportant n noeuds indicés par i. La matrice d'incidence du graphe, notée $\mathbf{X}$, est définie par $X_{ij} = 1$ s'il y a un arc du noeud i vers le noeud j et 0 sinon. On suppose que pour i, $X_{ii} = 0$.

Chaque noeud appartient à une classe q parmi Q classes possibles ($q \in \{1,\dots,Q\}$). À chaque noeud i est associée un vecteur $Z_i = (Z_{i1},\dots,Z_{iQ}) \in \{0,1\}^Q$ de *variables latentes*, indiquant l'appartenant du noeud i à la classe q : $Z_{iq} = 1$, si le noeud i appartient à la classe q, et 0 sinon. On introduit alors la matrice $\mathbf{Z} = \left(Z_{iq}\right)_{i,q}$, contenant les variables latentes discrètes d'appartenance aux classes.

Le modèle probabiliste est le suivant :

- $Z_i = (Z_{i1}, Z_{i2}...Z_{iQ}) \sim \mathcal{M}(1,\alpha_1,\alpha_2,\dots,\alpha_Q)$, où $\mathcal{M}(1,\alpha_1,\alpha_2,\dots,\alpha_Q)$ désigne la loi multinomiale de paramètres $(\alpha_1,\alpha_2,...\alpha_Q)$. Pour tout q, $0 < \alpha_q < 1$ représente la probabilité pour un noeud i d'appartenir à la classe q.

- Conditionnellement à $\mathbf{Z}$, les variables X_{ij} sont des variables *indépendantes* de loi de Bernoulli donnée par
$$P(X_{ij} = 1 \mid Z_{iq} = 1, Z_{jl} = 1) = \pi_{ql}.$$

Ce modèle permet de prendre en compte différentes structures topologiques : modules séparés, structure hiérarchique, *hubs*. Il permet également d'analyser des graphes orientés on non orientés (voir [2] et [4]).

2.1 Estimation des paramètres

La *vraisemblance complète* est donnée par:

$$\mathscr{L}(\mathbf{X},\mathbf{Z}) = \sum_i \sum_q \mathrm{Z}_{iq} \ln \alpha_q + \sum_{i<j} \sum_{q,l} \mathrm{Z}_{iq}\mathrm{Z}_{jl} \ln b(\pi_{ql}, \mathrm{X}_{ij})$$

où $b(\pi_{ql}, \mathrm{X}_{ij}) = \pi_{ql}^{\mathrm{X}_{ij}} (1-\pi_{ql})^{(1-\mathrm{X}_{ij})}$. La vraisemblance est définie à partir de la vraisemblance complète :

$$\mathscr{L}(\mathbf{X}) = \ln \sum_{\mathbf{Z}} \exp \mathscr{L}(\mathbf{X},\mathbf{Z})$$

Le calcul de la vraisemblance est impossible dès que $n > 30$, car il demande de calculer une somme de Q^n termes. De plus, l'algorithme EM repose sur le calcul de $\Pr(\mathbf{Z}|\mathbf{X})$, ce qui est impossible du fait de la dépendance (conditionnelle à X) des Z_i. On utilise donc une méthode approchée pour estimer les paramètres, l'*inférence variationnelle* qui consiste à remplacer la vraisemblance par une quantité calculable :

$$\begin{aligned} \mathscr{J}(\mathscr{R}_{\mathbf{X}}[\mathbf{Z}]) &= \mathscr{L}(\mathbf{X}) - \mathrm{KL}(\mathscr{R}_{\mathbf{X}}[\mathbf{Z}], \Pr(\mathbf{Z}|\mathbf{X})) \\ &= \sum_{i,j\neq i} \sum_{ql} b_{ij}(q,l)\tau_{iq}\tau_{jl} - \sum_{iq} \tau_{iq} \log \tau_{iq} + \sum_{iq} \tau_{iq} \log \alpha_q \end{aligned}$$

où $\mathrm{KL}(\cdot,\cdot)$ désigne l'information de Kullback-Leibler, et les paramètres variationnels τ_{iq} vérifient les contraintes $\tau_{iq} > 0$ et $\forall i = 1,\ldots,n$, $\sum_q \tau_{iq} = 1$.

À notre connaissance, il n'y a aucune propriété générale démontrée à ce jour pour les estimateurs variationnels ([5]). Seules certaines situations spécifiques ont été explorées. Pour certaines d'entre elles, les estimateurs variationnels sont consistants ([6]), mais pas dans d'autres ([7]).

3 Résultats

3.1 Modèle semi-paramétrique

On considère le modèle Mixnet conditionnellement aux variables latentes $\mathbf{Z} = z^*$. Ce modèle est appelé *Mixnet-FP*. Les paramètres de ce modèle sont z^* ($n(\mathrm{Q}-1)$ paramètres), et la matrice π (Q^2 paramètres).

Ce modèle est dit *semi-paramétrique* puisqu'à chaque noeud correspondent plusieurs paramètres à estimer et que le nombre total de paramètres du modèle croît avec n. L'estimation de ces paramètres est toutefois rendue possible par le nombre $n(n-1)$ d'observations disponibles pour chacun d'eux.

Résultat 1
Si $\forall (q \neq q')\ \ \exists l \in \{1,\ldots,\mathrm{Q}\}\quad \pi_{ql} \neq \pi_{q'l}$ ou $\pi_{lq} \neq \pi_{lq'}$, le modèle Mixnet-FP est identifiable et les estimateurs du maximum de vraisemblance de (π, z^*) sont consistants.

3.2 Modèle Mixnet

Plaçons nous à présent dans le cadre du modèle Mixnet, où les variables Z_i d'appartenance aux classes sont générées suivant la loi multinomiale de paramètres $(\alpha_1,\ldots,\alpha_Q)$. Dans ce modèle, les paramètres à estimer sont les α_q (Q paramètres), ainsi que la matrice π (Q^2 paramètres).

Résultat 2
Si le modèle Mixnet est identifiable, les estimateurs variationnels de (α,π) sont consistants et asymptotiquement équivalents aux estimateurs du maximum de vraisemblance.

Bibliographie

[1] A. W. van der Vaart and J. A. Wellner (1996), Weak Convergence and Empirical Processes With Applications to Statistics,Springer Series in Statistics.
[2] J.-J. Daudin & F. Picard & S. Robin (2008), A mixture model for random graphs,Stat. Comput., 18, 2008,173–183
[3] K. Nowicki & T. Snijders (2001),Estimation and prediction for stochastic block-structures, J. Am. Stat. Assoc., 96,1077–1087.
[4] F. Picard & V. Miele & J.-J. Daudin & L. Cottret & S. Robin (2009), Deciphering the connectivity structure of biological networks using MixNet,BMC Bioinformatics, 10,
[5] Asela Gunawardana & William Byrne (2005), Convergence Theorems for Generalized Alternating Minimization Procedures, JMLR
[6] P Hall & K Humphreys & DM Titterington (2002), On the adequacy of variational lower bound functions for likelihood-based inference in Markovian models with missing values,JRSSB, 64(3), 549–564.
[7] B. Wang & D.M. Titterington (2004), Lack of consistency of mean field and variational Bayes approximations for state space models, Neural Proceeding Letters, 20(3), 151–170.
[8] P. J. Bickel , A. Chen (2009) A nonparametric view of network models and Newman-Girvan and other modularities, PNAS