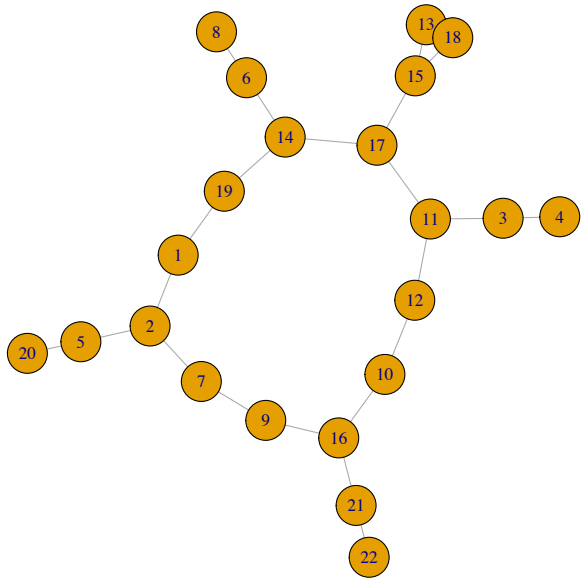
Analyse statistique de graphes

Chapitre 3: Modèles de graphes aléatoires et clustering des nœuds

Tabea Rebafka

2019

Master de statistique



SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

LPSM
LABORATOIRE DE PROBABILITÉS STATISTIQUE & MODÉLISATION

Outline

Modèles de graphes aléatoires

Déjà vu :

- Modèle $G(n, p)$
  - distribution uniforme des arêtes
  - pas de hubs
  - s'ajuste mal aux graphes observés.
- Modèle de configuration et loi de puissance
  - difficultés de simulation
  - les degrés des nœuds ne décrivent pas bien la complexité d'un réseau.

Modèle de graphe exponentiel I

Exponential Random Graph Model (ERGM)

- Notons $\mathcal{A}_n$ l'ensemble des matrices d'adjacence binaires (symétriques ou non) de taille $n \times n$.
- Pour tout $A \in \mathcal{A}_n$, soit $S(A) \in \mathbb{R}^p$ un vecteur de statistiques du graphe associé.
- Le **modèle exponentiel de graphe** associé à S, noté ERGM(S), est défini par la famille de lois de probabilités $\{\mathbb{P}_\theta\}_{\theta \in \mathbb{R}^p}$ définies sur l'ensemble $\mathcal{A}_n$ par

$$\forall \theta \in \mathbb{R}^p, \forall A \in \mathcal{A}_n, \quad \mathbb{P}_\theta(A) = \frac{1}{c(\theta)} \exp\left(\theta^\intercal S(A)\right),$$

avec $c(\theta) = \sum_{A \in \mathcal{A}_n} \exp(\theta^\intercal S(A))$ la constante de normalisation.

Modèle de graphe exponentiel II

- Par construction, l'ERGM(S) est une famille exponentielle avec S comme statistique exhaustive.
- $S(A)$ peut contenir
  - la taille du graphe, les degrés, le nombre de triangles, de k-cliques, de k-stars, ...
  - ou encore des **covariables** du modèle.

Modèle de graphe exponentiel III

Exemple 1

- Soit $S(A) = \mathrm{vec}(A) = \mathrm{vec}((A_{ij})_{1 \leq i < j \leq n})$.
- Alors le ERGM(S) correspondant vérifie

$$\mathbb{P}_\theta(A) \propto \exp\left\{\sum_{i<j} \theta_{ij} A_{ij}\right\} = \prod_{i<j} \exp\left\{\theta_{ij} A_{ij}\right\},$$

- Les A_{ij} sont alors indépendantes de loi Bernoulli $A_{i,j} \sim \mathcal{B}(p_{ij})$ avec $p_{ij} = \exp(\theta_{ij})/(1 + \exp(\theta_{ij}))$.
- Il y a autant de paramètres que d'observations, donc pas très pratique.
- Avec la contrainte $\theta_{ij} = \theta$ pour tout i, j, on obtient le modèle d'Erdös-Rényi.

Modèle de graphe exponentiel IV

Exemple 2

- Soit $S(A) = (\sum_{i,j} A_{ij}, \sum_{i,j,k} A_{ij}A_{ik})$ le vecteur de la taille du graphe et du nombre de chemins de longueur 2.
- Alors les variables $(A_{ij})_{i<j}$ sont non indépendantes et on n'a pas d'expression analytique pour l'EMV.

Exemple 3

- Notons $k \geq 1$ et $S_k(A)$ le nombre de k-stars du graphe A et $T(A) = \sum_{ijk} A_{ij}A_{ik}A_{jk}$ le nombre de triangles.
- Dans le **Markov random graph**, on utilise $S = (S_1, \ldots, S_{n-1}, T)$.
- En pratique, aller jusqu'à $k = n - 1$ est beaucoup trop grand et on se contente de $k \ll n - 1$ pour la plupart des ERGM courants.

Modèle de graphe exponentiel V

Problèmes du ERGM

- La constante $c(\theta)$ n'est pas calculable. Les méthodes d'estimation sont basées sur des méthodes MCMC avec par exemple un échantillonneur de Gibbs pour supprimer le problème de la constante inconnue.
- La maximisation de la vraisemblance reste un problème difficile, et en fait mal posé : ces modèles sont souvent 'dégénérés' au sens où cette loi concentre sa masse sur le graphe complet ou le graphe vide, ou un mélange des deux. Voir Chatterjee and Diaconis (2013); Schweinberger and Handcock (2015) pour plus de détails.
- Problème important du bon choix du vecteur des statistiques $S(A)$.

Je déconseille fortement l'usage des ERGM.

Attachement préférentiel I

Modèle dynamique d'évolution des graphes selon le principe *rich get richer*.

Modèle par attachement préférentiel

- Initialisation : un petit graphe $G_0 = (V_0, E_0)$, le nombre m d'arêtes à ajouter à chaque itération, le nombre k_{final} d'itérations.
- Itération k :
  - Notons $G_{k-1} = (V_{k-1}, E_{k-1})$ le graphe actuel et la suite de degrés associés $(d_{k-1,1}, \ldots, d_{k-1,|V_k|})$.
  - On fabrique le graphe $G_k = (V_k, E_k)$ en ajoutant un nouveau nœud nœud v_k : $V_k = V_{k-1} \cup \{v_k\} = V_0 \cup \{v_1, \ldots, v_k\}$ et
  - en connectant ce nouveau nœud avec m nœuds existants qui sont choisis chacun avec probabilité $d_{j,k-1}/(2|E_{k-1}|)$.

Principe : attachement préférentiel aux nœuds de degrés les plus élevés.

Attachement préférentiel II

Avantages et inconvénients

- Le model est génératif dynamique.
- Il permet d'expliquer la loi de puissance des degrés : à la limite ($k \to \infty$) et sous certaines conditions, la distribution des degrés du graphe suit une loi de puissance.
- Problème du choix des paramètres G_0, m, k_{final}. Impact de ce choix sur le graphe obtenu ?
- D'un point de vue statistique, ce n'est pas un modèle qu'on peut ajuster sur les données.

Modèles à variables latentes : quelques généralités I

- Modèles graphiques
- Modèles à variables latentes
- Estimation dans le modèle à variables latentes
- Algorithme EM

Modèle à positions latentes I

- Modèle de graphe pour l'étude des réseaux sociaux
- Des variables latentes à valeurs dans $\mathbb{R}^q$ représentent un **espace social** (non observé).
- La proximité des individus dans cet espace induit une plus grande probabilité de connexion dans le graphe.

Modèle à positions latentes II

Modèle à positions latentes (Hoff et al., 2002)

- Soit $(A_{ij})_{1 \leq i,j \leq n}$ la matrice d'adjacence d'un graphe binaire non dirigé.
- Il est possible d'inclure des covariables $\mathbf{x}_{ij} \in \mathbb{R}^s$ sur les relations (i, j).
- Soient Z_i i.i.d. des variables latentes à valeurs dans $\mathbb{R}^q$.
- On considère un modèle de régression logistique

$$\begin{aligned} \text{logit}(\mathbb{P}(A_{ij} = 1 | Z_i, Z_j, \mathbf{x}_{ij})) &= \frac{\mathbb{P}(A_{ij} = 1 | Z_i, Z_j, \mathbf{x}_{ij})}{1 - \mathbb{P}(A_{ij} = 1 | Z_i, Z_j, \mathbf{x}_{ij})} \\ &= \alpha + \beta^\intercal \mathbf{x}_{ij} - \|Z_i - Z_j\|, \end{aligned}$$

  où $\|\cdot\|$ est la norme euclidienne dans l'espace latent $\mathbb{R}^q$ (ou n'importe quelle distance).
- Les paramètres du modèle sont $(\alpha, \beta) \in \mathbb{R} \times \mathbb{R}^s$.

Modèle à positions latentes III

- La probabilité de connexion ne dépend que de la position relative des variables latentes entres elles (et pas de leur position absolue).
- On appelle **configurations équivalentes** deux ensembles $\{Z_i\}_i$ et $\{Z'_i\}_i$ qui induisent les mêmes valeurs de distances $(\|Z_i - Z_j\|)_{i,j} = (\|Z'_i - Z'_j\|)_{i,j}$.
- Le paramètre α règle la densité du graphe.
- Estimation des paramètres et des variables latentes : avec le package `latentnet` (méthode bayésienne, voir TP).

Modèle à positions latentes IV

Version classifiante du modèle

- Pour obtenir un clustering des nœuds, on peut introduire un modèle de mélange sur les variables latentes (Handcock et al., 2007).
- On suppose que les $Z_i \in \mathbb{R}^q$ sont générées selon un mélange de lois gaussiennes multi-dimensionnelles $\mathcal{N}_q(m_k, \sigma_k^2 I)$ avec $1 \leq k \leq K$, de proportions $\pi_k, 1 \leq k \leq K$, de moyennes différentes $(m_k, 1 \leq k \leq K)$ et des matrices de covariance sphériques $(\sigma_k^2 I)$.
- Le choix du nombre de clusters K se fait automatiquement dans le cadre bayésien : on place une loi a priori sur K et on l'estime par le maximum a posteriori.
- Les groupes obtenus sont nécessairement des communautés : si deux variables Z_i, Z_j sont dans la même composante gaussienne, alors elles sont proches dans $\mathbb{R}^q$ et la probabilité que les nœuds i, j soient connectés est plus grande.

Modèle à positions latentes V

Choix de la dimension de l'espace latent

- En pratique, il n'existe aucun méthode permettant de choisir la dimension q de l'espace latent (attention, cette dimension n'est pas le nombre de clusters K de la méthode de Handcock et al. (2007) !).
- Les logiciels sont implémentés avec $q = 2$ (ou 3), mais rien ne permet d'affirmer que ce choix est pertinent, ni qu'il n'a pas un impact majeur sur les résultats.

Modèle à blocs stochastiques

- Définition
- Estimation des paramètres et clustering
- Approximation variationnelle
- Sélection du nombre de clusters

Références I

Chatterjee, S. and P. Diaconis (2013). Estimating and understanding exponential random graph models. *The Annals of Statistics 41*(5), 2428–2461.

Handcock, M., A. Raftery, and J. Tantrum (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society : Series A (Statistics in Society) 170*(2), 301–54.

Hoff, P., A. Raftery, and M. Handcock (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc. 97*(460), 1090–98.

Schweinberger, M. and M. S. Handcock (2015). Local dependence in random graph models : characterization, properties and statistical inference. *Journal of the Royal Statistical Society : Series B (Statistical Methodology) 77*(3), 647–676.