Modèles statistiques pour les graphes aléatoires hétérogènes, application aux réseaux biologiques et sociaux

Jean-Jacques Daudin

Modèles statistiques pour les graphes aléatoires hétérogènes, application aux réseaux biologiques et sociaux

Jean-Jacques Daudin

UMR AgroParisTech/INRA MIA 518, 16 rue C. Bernard Paris 5°

Mots-clés Biologie-génôme, Données économiques et sociales

Résumé

Les données qui sont sous la forme de mesures de relations entre items sont de plus en plus disponibles, comme en témoigne l'extraordinaire développement des réseaux sociaux et des réseaux biologiques. Ce type de données oblige le statisticien à abandonner la structure usuelle d'un jeu de données de type individus-variables pour une structure de type individu-individu. Ces données "relationnelles" sont très souvent présentées sous la forme d'un graphe, même si cette représentation a ses limites, notamment quand le nombre d'individus dépasse la centaine. Le modèle probabiliste de graphe uniforme étudié par Erdos est trop simple pour représenter correctement les réseaux réels. Il faut donc développer des modèles pour des graphes hétérogènes. Dans des travaux précédents nous avions développé un modèle de mélange basé sur des variables latentes discrètes, appelé MIXNET, pour lequel nous avions utilisé une méthode d'estimation variationnelle. Dans cet exposé, nous proposons un nouveau modèle, appelé "Ideal-Type model", dans lequel les variables latentes discrètes sont remplacées par des paramètres continus caractérisant chaque individu. Nous analysons l'identifiabilité de ce modèle et proposons un algorithme pour obtenir les estimateurs du maximum de vraisemblance. Pour illustrer la méthode, un réseau social et un réseau biologique sont analysés.

Summary

Data sets giving not only information about items but also information about the relation between them are more and more studied in different domains such as social sciences and biology. The data size is proportional to the square of the number of individuals, so that it is necessary to summarize the information in a simpler form. The network representation of the data is graphically attractive, but is not readable for $n > 100$. There are two ways for producing a synthetic representation of such data: multidimensional scaling where position in a metric space is assigned to each item, and clustering of the items using a mixture model. In this paper we present a new method, which has some flavor of mixture and some flavor of multidimensional scaling but is really different of both. We restrict out interest to the case of pure relational information, putting aside any information on items. The intensity of relation may be continuous or binary. We restrict our interest on the binary case. These two restrictions are made for sake of simplicity. The model

we propose may be extended to the general case, but this is not done in this paper. We define the Ideal-Type Model (IDTM), give a maximum-likelihood estimation algorithm and some examples from biological and social networks.

Le modèle IDTM

On considère un graphe avec n sommets, $\{1, \ldots, n\}$. Le modèle est basé sur Q sommets "Idéaux" non observés. Cette terminologie est empruntée au sociologue Max Weber, mais le modèle est parfaitement général.

Chaque sommet i est la moyenne pondérée des Q sommets idéaux, avec les poids donnés par $Z_i = (z_{i1}, \ldots, z_{iQ})$, avec $z_{iq} \geq 0$ and $\sum_q z_{iq} = 1$. Les Q sommets idéaux sont placés arbitrairement aux points terminaux des vecteurs unitaires canoniques $(1, 0...0)$, $(0, 1, 0...0)...(0...0, 1)$ de $\mathbb{R}^Q$. Les sommets $\{1, \ldots, n\}$ appartiennent au simplexe $S_Q = \{x, \in [0, 1]^Q, \sum_{q=1,Q} x_q = 1\}$, dont les sommets idéaux sont les sommets extrêmes.

L' arc du sommet i au sommet j est associé à une variable aléatoire binaire X_{ij} qui a une distribution de Bernoulli de probabilité P_{ij}. La probabilité que le sommet idéal q envoie un arc vers le sommet idéal l est a_{ql} et P_{ij} est défini par

$$P_{ij} = \sum_{q,l=1,Q} z_{iq} a_{ql} z_{jl}$$

soit

$$P = ZAZ',$$

avec

- P la matrice (n, n) contenant les p_{ij},
- Z la matrice (n, Q) contenant les z_{iq} and Z' la transposée de Z, $Z \in S_Q^n$,
- $A \in [0, 1]^{Q^2}$, la matrice (Q, Q) contenant les a_{ql}, matrice de connectivité entre les idéaux.

Les X_{ij} sont supposées independantes. Soit X la matrice (n, n) contenant les X_{ij}. Le modèle est

$$X \sim \mathbb{B}(Z'AZ)$$

où $\mathbb{B}$ est la loi de Bernoulli, $Z \in S_Q^n$ et $A \in [0, 1]^{Q^2}$.

Les paramètres du modèle sont A et Z. En un certain sens on pourrait considérer que c'est un modèle semi-paramétrique parce que chaque individu a son propre jeu de paramètres $(z_{i1}, ...z_{iQ})$. En général en statistique il est impossible d'estimer autant de paramètres que d'individus. De plus il y a $Q^2 + n(Q - 1)$ paramètres, nombre qui tend vers l'infini avec n. Cependant le nombre d'observations contenues dans X n'est pas proportionel à n mais à n^2, de telle sorte que le rapport du nombre de paramètres sur le nombre d'observations tends vers 0 quand $n \to \infty$. En pratique pour chaque sommet il y a n données, $(x_{i1}, ...x_{in})$, pour estimer $Q - 1$ parameters $(z_{i1}, ...z_{iQ})$.

Le graphe peut être orienté (X quelconque) ou non (X symétrique) On suppose qu'il n'y a pas de boucle individuelle ($X_{ii} = 0$, pour $i = 1, n$).

Identifiabilité Le modèle n'est pas identifiable mais on peut le rendre identifiable en ajoutant une condition.

Estimation
La log-vraisemblance :

$$L = \sum_{i,j} x_{ij} \log(\sum_{q,l=1,Q} z_{iq} A_{ql} z_{jl}) + (1 - x_{ij}) \log(1 - \sum_{q,l=1,Q} z_{iq} A_{ql} z_{jl}) \quad (1)$$

$$L(A, Z) = Tr(X' \log(ZAZ')) + Tr((J - X)' \log((J - ZAZ')))$$

où J est une matrice (n, n) composée de 1, et les contraintes sur les paramètres sont

$$\begin{aligned} A &\in [O, 1]^{n^2} \\ Z &\in S_Q^n \end{aligned}$$

Dérivées de la log-vraisemblance

$$\frac{\partial L}{\partial Z} = RZA' + R'ZA$$

où R est une matrice (n, n) avec $r_{ij} = \frac{x_{ij} - p_{ij}}{p_{ij}(1-p_{ij})}$, et

$$\frac{\partial L}{\partial A} = Z'RZ$$

$$\frac{\partial L}{\partial a_{ql} \partial a_{uv}} = -\sum_{ij} r_{ij}^2 z_{iq} z_{iu} z_{jl} z_{jv}$$

$$\frac{\partial L}{\partial a_{ql} \partial z_{iu}} = \delta_{qu} \sum_j r_{ij} z_{jl} + \delta_{lu} \sum_j r_{ji} z_{jq} - \sum_{jv} (r_{ij}^2 z_{jl} z_{jv} a_{uv} z_{iq} + r_{ji}^2 z_{jq} z_{jv} a_{vu} z_{jq})$$

$$\frac{\partial L}{\partial z_{iq} \partial z_{jl}} = r_{ij} a_{ql} + r_{ji} a_{lq} - \delta_{ij} \sum_{kuv} \left[r_{ik}^2 z_{ku} a_{lu} z_{kv} a_{qv} + r_{ki}^2 z_{ku} a_{ul} z_{kv} a_{vq} \right] - \sum_{uv} \left(r_{ij}^2 z_{iu} a_{ul} z_{jv} a_{qv} + r_{ji}^2 z_{iu} a_{lu} z_{jv} a_{vq} \right)$$

où $\delta_{ij} = 1$ si $i = j$, et $\delta_{ij} = 0$ si $i \neq j$.

Algorithme Les contraintes sur les paramètres sont linéaires, mais la log-likelihood ne l'est pas. La linéarisation de (1) conduit à une problème de programmation linéaire.

Soit $A^{(k)}$ et $Z^{(k)}$ les valeurs des paramètres estimés à l'étape k, $P^{(k)} = Z^{(k)} A^{(k)} Z^{(k')}$ et $R^{(k)}$ telle que $r_{ij}^{(k)} = \frac{x_{ij} - p_{ij}^{(k)}}{p_{ij}^{(k)}(1-p_{ij}^{(k)})}$.

L'approximation linéaire de la log-vraisemblance au point $(A^{(k)}, Z^{(k)})$ est

$$\begin{aligned} L(A,Z) &\approx L(A^{(k)},Z^{(k)}) + Tr\left[(A-A^{(k)})'\frac{\partial L}{\partial A}(A^{(k)},Z^{(k)})\right] + Tr\left[(Z-Z^{(k)})'\frac{\partial L}{\partial Z}(A^{(k)},Z^{(k)})\right] \\ &\approx L(A^{(k)},Z^{(k)}) + Tr\left[(A-A^{(k)})'Z^{(k)\prime}R^{(k)}Z^{(k)}\right] \\ &+ Tr\left[(Z-Z^{(k)})'(R^{(k)}Z^{(k)}A^{(k)\prime} + R^{(k)\prime}Z^{(k)}A^{(k)})\right] \end{aligned}$$

D'où l'algorithme

- Valeurs initiales $(A^{(0)}, Z^{(0)})$
- A l'étape (k) on utilise la programmation linéaire pour maximiser en (A, Z) la fonction

  $$f_k(A,Z) = Tr\left[A'Z^{(k)\prime}R^{(k)}Z^{(k)}\right] + Tr\left[Z'(R^{(k)}Z^{(k)}A^{(k)\prime} + R^{(k)\prime}Z^{(k)}A^{(k)})\right]$$

  sous les contraintes

  $$\begin{aligned} A &\in [O,1]^{n^2} \\ Z &\in S_Q^n \end{aligned}$$

- Règles d'arrêt

  $$\begin{aligned} &\|A^{(k)} - A^{(k-1)}\| + \|Z^{(k)} - Z^{(k-1)}\| < \\ &|L(A^{(k)},Z^{(k)}) - L(A^{(k-1)},Z^{(k-1)})| < \alpha \end{aligned}$$

In pratique on borne les différences entre 2 itérations successives en ajoutant les contraintes $|Z^{(k)} - Z^{(k-1)}| < \epsilon_k$ and $|A^{(k)} - A^{(k-1)}| < \epsilon_k$, avec ϵ_k qui décroit avec k. Par ailleurs on calcule la log-vraisemblance en plusieurs points du segment reliant $Z^{(k)}, A^{(k)}$ et $Z^{(k-1)}, A^{(k-1)}$, ce qui permet de prendre en compte la non-linéarité de la fonction à optimiser.

Exemples
Un exemple de réseau social (Zachary et al, 1977) et un exemple de réseau biologique.

Conclusions
Un modèle de mélange est une façon souple de modéliser un graphe aléatoire hétérogène (Nowicki et al. 2001 et Daudin et al., 2007). Elle permet de prendre en compte la plupart des structures topologiques : structure de communauté, structure hiérarchique, hubs... Une faiblesse de ce modèle provient du fait qu'il n'existe pas de méthode réellement satisfaisante du point de vue théorique pour estimer les paramètres. La méthode dite

"EM variationnel" (Daudin et al. 2007) n'est pas toujours consistante et ne permet pas de récuperer les variances asymptotiques des estimateurs. Par ailleurs, les méthodes MCMC se heurtent au défi de la dimension de l'espace à explorer, $(Q-1)^n$, qui plus est avec des variables discrètes, ce qui fait qu'elles ne permettent pas d'analyser des graphes de plus de 200 sommets. Le modèle IDT est une réponse à cet obstacle. L'astuce consiste à remplacer l'espace discret par un espace continu, ce qui améliore considérablement les conditions de l'optimisation et à se plonger dans un espace de contraintes linéaires, ce qui permet d'utiliser un algorithme efficace, la programmation linéaire. Il reste du travail pour comprendre parfaitement le comportement asymptotique d'un modèle comportant un nombre de paramètres qui augmente linéairement avec n et un nombre d'observations qui est quadratique en n.

Bibliographie

[1] Daudin, JJ., Picard, F. and Robin, S. (2007). A mixture model for random graphs. *Statis. Comput.* **18(2)**, 173–183

[2] Handcock, MS., Raftery, AE. and Tantrum, JM. (2007). Model-based clustering for social networks. *JRSSA* **54**, 301–354

[3] Nowicki, K. and Snijders, T. (2001). Estimation and prediction for stochastic block-structures. *J. Am. Stat. Assoc.* **96** ,1077–1087

[4] Zachary, WW. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research* **33**, 452–473