Modèles à blocs latents pour graphe multipartite

Application aux interactions entre espèces animales et plantes.

Sophie Donnet [1], Avner Bar-Hen [2] & Pierre Barbillon [3]

[1] *UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France, sophie.donnet@agroparistech.fr*
[2] *CNAM, avner@cnam.fr*
[3] *UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France, pierre.barbillon@agroparistech.fr*

Résumé. Dans l'environnement naturel, une grande diversité des types d'interactions entre plantes et espèces animales (parmi lesquels la pollinisation, la dissémination de graines, etc) coexiste. Chaque type d'interaction peut être représenté par un graphe bipartite entre l'ensemble de plantes observé et un groupe fontionnel d'animaux donné. Jusqu'à récemment, les structurations de ces différents réseaux d'interactions étaient étudiées séparément. Dans ce travail, nous proposons de modéliser de façon conjointe les différents types d'interactions au moyen d'une extension adéquate des modèles à blocs latents. Nous utilisons une version variationnelle de l'algorithme EM pour maximiser la vraisemblance du modèle et développons un critère pénalisé de sélection de modèle adapté au problème. La pertinence des méthodes et du modèle est illustrée sur données simulées et réelles.

Mots-clés. Graphe multipartite, Modèle à blocs latents, EM variationnel, Écologie

Abstract. Within the natural environment, there is a high diversity of types of interactions between plants and animal species (namely pollinisation, seed dispersal, etc). Any type of interaction can be represented as a bipartite graph between the plants and a given group of animal species. Until recently, the structures of such networks were studied separately. In this work, we propose a joint modelisation of the various types of interactions thanks to an extension of the latent blocks model. The likelihood function of the model is maximized with a variational version of the EM algorithm. We develop an adapted penalized model selection criterion. The model and the procedure are tested on simulated and real datasets.

Keywords. Multipartite graph, Latent block models, Variational EM, Ecology

1 Problématique

Dans tout environnement naturel, il existe une grande diversité des interactions entre plantes et espèces animales. Ces interactions écologiques, qui peuvent être du type pollinisation, dissémination de graines, etc. régulent les populations et jouent un rôle déterminant dans la structuration de la biodiversité. Récemment, les outils de l'analyse de réseaux ont été utilisés pour comprendre l'organisation complexe des réseaux d'interaction entre espèces. Alors qu'auparavant, les différents types d'interactions étaient étudiés et analysés séparement, des travaux considérant conjointement ces différents types d'interaction ont été récemment proposés [Dáttilo et al., 2016, Fontaine et al., 2011, Kéfi et al., 2016].

Si l'on considère un unique groupe fonctionnel d'espèces animales (par exemple pollinisateurs, papillons, fourmis, animaux ou disséminateurs de graines etc.), correspondant à un unique type d'interaction, l'ensemble des relations peut être représenté par un graphe bipartite entre l'ensemble de plantes observées et ce groupe fonctionnel donné. À ce graphe correspond une matrice d'adjacence, avec en ligne les plantes et en colonne les différentes espèces animales de ce groupe fonctionnel. Parmi les outils disponibles dans la littérature, les modèles à blocs latents fournissent un cadre probabiliste pour la classification simultanée des lignes (ici plantes) et des colonnes (ici espèces animales) de la matrice [Govaert and Nadif, 2008]. Les espèces (animales ou végétales) ainsi classées ensemble partagent un même comportement de connexion.

Dans ce travail, nous nous intéressons à l'extension des modèles stochastiques à blocs latents aux graphes multipartites dans lesquels on considère tous les types d'interactions entre les plantes et les groupes fonctionnels. Nous cherchons donc à la fois une classification des plantes prenant en compte tous les types d'interactions mais aussi une sous-classification de chaque groupe fonctionnel d'espèces animales.

2 Modèle et inférence

Considérons un groupe de n_0 plantes et Q groupes fonctionnels d'espèces animales comptant respectivement $n_1, \dots n_Q$ espèces. Pour chaque groupe fonctionnel q, nous notons $X^q \in \{0,1\}^{n_0 n_q}$ la matrice d'adjacence entre les plantes et les espèces animales du groupe q, i.e. $X^q_{ij} = 1$ si l'espèce j du groupe q a été observée directement sur la plante i. Afin de modéliser l'hétérogénéité des relations, nous introduisons la structure latente suivante. Supposons que les plantes sont réparties en K_0 classes, et chaque groupe fonctionnel q est scindé en K_q classes ($q = 1 \dots Q$). Pour tout $q = 0 \dots Q$, nous notons $(Z^q_j)_{j=1 \dots n_q}$ les variables aléatoires d'affectations aux classes. Elles sont supposées indépendantes et distribuées de la façon suivante :

$$P(Z^q_i = k) = \pi^q_k, \quad \forall k = 1 \dots K_q, \forall i = 1 \dots n_q, \forall q = 0 \dots Q \tag{1}$$

avec $\sum_{k=1}^{K_q} \pi^q_k = 1$ pour tout $q = 0, \dots Q$.

Conditionnellement aux variables latentes $\boldsymbol{Z} = \{Z_i^q, i = 1 \dots n_q, q = 0 \dots Q\}$, les observations $\boldsymbol{X} = \{X^q, q = 1 \dots Q\}$ sont modélisées de la façon suivante :

$$X_{ij}^q | Z_i^0, Z_j^q \sim_{i.i.d} \mathcal{B}ern(\alpha_{Z_i^0, Z_j^q}) \, .$$

Remarque 1 : Soulignons que la dépendance entre les différentes interactions (codées par les matrices d'adjacences) est introduite par le fait que toutes les probabilités de connexion dépendent des variables d'affectation des plantes $\boldsymbol{Z}^0 = (Z_1^0, \dots, Z_{n_0}^0)$.

Remarque 2 : Notre modèle revient à un modèle à blocs latents pour un graphe bipartite entre les plantes et l'ensemble des espèces animales dans lequel les classes des espèces animales seraient contraintes par les groupes fonctionnels. Plus précisément, une classe d'espèces animales doit être une sous partie d'un groupe fonctionnel connu.

Les paramètres d'intérêt sont les probabilités de connexion

$$\boldsymbol{\alpha} = \{\alpha_{kk'}, k = 1 \dots K_0, k' = 1 \dots K_q, q = 1 \dots Q\}$$

et les probabilités d'affections aux groupes

$$\boldsymbol{\pi} = \{\pi_k^q, k = 1 \dots K_q, q = 0 \dots Q\} \, .$$

La vraisemblance des données dites complètes i.e. $(\boldsymbol{X}, \boldsymbol{Z})$ s'écrit alors :

$$\mathcal{L}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\pi}, \boldsymbol{\alpha}) = \prod_{q=1}^{Q} \prod_{i=1}^{n_0} \prod_{j=1}^{n_q} \alpha_{Z_i^0, Z_j^q}^{X_{ij}^q} (1 - \alpha_{Z_i^0, Z_j^q})^{1 - X_{ij}^q} \prod_{q=0}^{Q} \prod_{i=1}^{n_q} \pi_{Z_i^q}^q \, . \tag{2}$$

La vraisemblance des données observées $\boldsymbol{X}$ est obtenue en sommant l'expression précédente (2) sur toutes les valeurs possibles d'affectations aux classes $\boldsymbol{Z}$. Cette intégration devient rapidement impossible à effectuer d'un point de vue calculatoire quand les nombres de classes $K_0, K_1, \dots, K_Q$ augmentent. Dans ce cas, l'algorithme Expectation Maximisation dans sa version variationnelle est un outil puissant de maximisation de la vraisemblance [Govaert and Nadif, 2008]. Nous l'appliquons à ce modèle.

Concernant le choix des nombres de classes $K_0 \dots, K_Q$, nous avons recours au critère pénalisé ICL (Integrated Classification Likelihood) proposé par [Biernacki et al., 2000] et adapté au cas des modèles à blocs latents standards par [Keribin et al., 2014]. Nous l'étendons ici à notre modèle à blocs latents pour la modélisation conjointe de plusieurs types d'interactions.

Nous illustrons la robustesse de notre procédure d'inférence sur des jeux de données simulées et illustrons la pertinence de notre modèle sur un jeu de données réelles issu d'un environnement côtier tropical du Mexique [Dáttilo et al., 2016].

Références

[Biernacki et al., 2000] Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(7) :719–725.

[Dáttilo et al., 2016] Dáttilo, W., Lara-Rodríguez, N., Jordano, P., Guimarães, P. R., Thompson, J. N., Marquis, R. J., Medeiros, L. P., Ortiz-Pulido, R., Marcos-García, M. A., and Rico-Gray, V. (2016). Unravelling darwin's entangled bank : architecture and robustness of mutualistic networks with multiple interaction types. *Proceedings of the Royal Society of London B : Biological Sciences*, 283(1843).

[Fontaine et al., 2011] Fontaine, C., Guimarães, P. R., Kéfi, S., Loeuille, N., Memmott, J., van der Putten, W. H., van Veen, F. J. F., and Thébault, E. (2011). The ecological and evolutionary implications of merging different types of networks. *Ecology Letters*, 14(11) :1170–1181.

[Govaert and Nadif, 2008] Govaert, G. and Nadif, M. (2008). Block clustering with bernoulli mixture models : Comparison of different approaches. *Comput. Stat. Data Anal.*, 52(6) :3233–3245.

[Keribin et al., 2014] Keribin, C., Brault, V., Celeux, G., and Govaert, G. (2014). Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, pages 1–16.

[Kéfi et al., 2016] Kéfi, S., Miele, V., Wieters, E. A., Navarrete, S. A., and Berlow, E. L. (2016). How structured is the entangled bank ? the surprisingly simple organization of multiplex ecological networks leads to increased persistence and resilience. *PLOS Biology*, 14(8) :1–21.