Inférence de réseaux à partir de mélanges d'arbres

Encadré par S. Robin[1] et C. Ambroise[12]

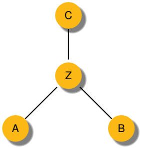
Raphaëlle Momal-Leisenring

[1] UMR AgroParisTech / INRA MIA-Paris
[2] LaMME, Evry

23 mars 2018

Réseau

- Réseau : représentation graphique de la structure de dépendance conditionnelle d'un jeu de données.
- Inférer un réseau : inférer les arêtes du graph, *i.e.* la structure de dépendance.



$\Rightarrow$

Les variables A, B et C sont inépendantes entre elles conditionnellement à la variable Z.

Exemple de réseau écologique

[**?**] :

- But : identifier les liens de dépendance entre le champignon *E. alphitoïde* présent sur les feuilles du chêne, et les autres micro-organismes présents.

- Utile à la compréhension et au contrôle des maladies chez le chêne.

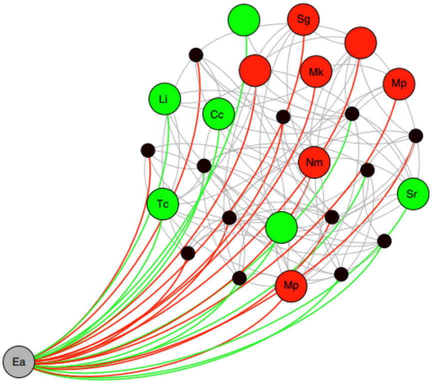


FIGURE – Model of the pathobiome *Erysiphe alphitoides* on oak leaves, source : Jakuschkin *et al.*

Modèle graphique

- Clique C d'un graphe G : sous-ensemble de noeuds de G qui sont tous liés entre eux.
- Clique maximale C_G : aucune autre clique de G ne la contient strictement.

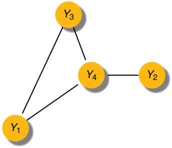
Propriété modèle graphique [?]

Soit $Y = (Y_1, ..., Y_q)$ et p sa densité. p se factorise selon le graphe non orienté G si :

$$p(y) \propto \prod_{C \in C_G} \Phi_C(y^C)$$

Et alors G représente la structure d'indépendance conditionnelle entre les Y_i.

Exemple : $Y = (Y_1, ..., Y_4)$:



$$p(Y) = \phi_1(Y_1, Y_4) \times \phi_2(Y_2, Y_3, Y_4)$$

Cas gaussien (GGM, *Gaussian Graphical Models*)

Soit Y une variable gaussienne multivariée de dimension d :

$$Y = (Y_1, ..., Y_d) \sim \mathcal{N}_d(0, \Sigma),$$
$$\Omega = \Sigma^{-1} = (w_{ij})_{1 \leq i,j \leq d}.$$

L'écriture de la gaussienne permet directement d'obtenir une factorisation :

$$\begin{aligned} p(y) &\propto exp(-y^T \Omega y/2) \\ &\propto \prod_{j,k,\omega_{jk} \neq 0} exp(-y_j w_{jk} y_k/2) \end{aligned}$$

Cas gaussien (GGM, *Gaussian Graphical Models*)

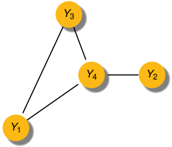
Soit Y une variable gaussienne multivariée de dimension d :

$$Y = (Y_1, ..., Y_d) \sim \mathcal{N}_d(0, \Sigma),$$
$$\Omega = \Sigma^{-1} = (w_{ij})_{1 \leq i,j \leq d}.$$

L'écriture de la gaussienne permet directement d'obtenir une factorisation :

$$\begin{aligned} p(y) &\propto exp(-y^T \Omega y/2) \\ &\propto \prod_{j,k,\omega_{jk} \neq 0} exp(-y_j w_{jk} y_k/2) \end{aligned}$$

$$\Omega = \begin{pmatrix} * & 0 & * & * \\ 0 & * & 0 & * \\ * & 0 & * & * \\ * & * & * & * \end{pmatrix} \quad \Rightarrow$$



Inférence de Ω : le graphical Lasso

- Estimation parcimonieuse
- La log-vraisemblance de Y s'écrit :

$$L(Y, \Omega) \propto \frac{n}{2} \log(det(\Omega)) - \frac{n}{2} Y^T \Omega Y$$

Le graphical-Lasso (glasso) :

Le graphical-Lasso pénalise la norme l_1 de la matrice de précision :

$$\widehat{\Omega}_\lambda = \arg \min_{\Omega \in \mathcal{S}_d^+} \left\{ L(Y, \Omega) + \lambda \sum_{i \neq j} |w_{ij}| \right\}$$

- Choix du λ...

Notation

- Y : tableau de données, de dimension $n \times d$
- d : nombre de variables (ex : espèces)
- n : nombre d'observations (échantillons)

Arbre de dépendance

- La structure de dépendance des données s'appuie sur un arbre
- Vraisemblance de données continues [**?**] :

$$\begin{aligned}
\mathbb{P}(Y|T) &= \mathbb{P}(Y_1|T)\prod_{j=2}^{d}\frac{\mathbb{P}(Y_j, Y_{a_j}|T)}{\mathbb{P}(Y_{a_j}|T)} \\
&= \underbrace{\prod_{j=1}^{d}\mathbb{P}(Y_j|T)}_{A}\prod_{(k,l)\in T}\underbrace{\frac{\mathbb{P}(Y_k, Y_l|T)}{\mathbb{P}(Y_k|T)\times\mathbb{P}(Y_l|T)}}_{\psi_{kl}(Y)} \\
&= A\prod_{k,l\in T}\psi_{kl}(Y)
\end{aligned}$$

- Cas gaussien centré réduit :

$$\log(\mathbb{P}(Y|T)) \propto \sum_{k,l\in T}\underbrace{-\frac{n}{2}\log(1-\hat{\rho}_{kl}^2)}_{\log(\hat{\psi}_{kl})} + \sum_{k}\underbrace{-\frac{n}{2}\log(\hat{\sigma}_k^2)}_{\log(\hat{A})}$$

Mélange d'arbres

En fait un mélange sur les arêtes du graph :

- Un poids β_{kl} est attribué à chaque arête (k, l) possible du graph.
- La probabilité de l'arbre de dépendance des données s'écrit alors

$$\mathbb{P}(T) = \frac{1}{B} \prod_{k,l \in T} \beta_{kl} \text{ , avec } B = \sum_{T \in \mathcal{T}} \prod_{k,l \in T} \beta_{kl}$$

- Les poids sont en général fixés à l'avance

L'idée

L'arbre réel qui structure la dépendance est caché.

$$\Downarrow$$

Construire un algorithme EM avec pour paramètre la loi de l'arbre, avec mise à jour des poids des arêtes.

Algorithme EM : étape E

$$\boxed{\log(p_\theta(Y)) = \mathbb{E}_\theta\left(\log(p_\theta(Y,Z))|Y\right) \underbrace{-\mathbb{E}_\theta\left(\log(p_\theta(Y|Z))|Y\right)}_{\mathcal{H}(p_\theta(Y|Z))}}$$

$$\mathbb{P}(Y,T) = \mathbb{P}(T) \times \mathbb{P}(Y|T)$$

$$\begin{aligned} \log(\mathbb{P}(Y,T)) &= \sum_{(k,l)\in E_T} [\log(\beta_{kl}) + \log(\psi_{kl})] - \log(B) + \log(A) \\ &= \sum_{k,l\in V} \mathbb{1}_{\{(k,l)\in E_T\}}(\log(\beta_{kl}) + \log(\psi_{kl})) - \log(B) + \log(A) \end{aligned}$$

Espérance conditionnelle :

$$\mathbb{E}_\theta[\log(\mathbb{P}(Y,T))|Y] = \sum_{k,l\in V} \mathbb{P}((k,l)\in E_T|Y)\,(\log(\beta_{kl}) + \log(\psi_{kl})) - \log(B) + \log(A)$$

Calcul de la probabilité conditionnelle

Théorème du matrix tree (étendu aux réels, [?])

Pour toute matrice symétrique $W = (a_{kl})_{k,l}$, son Laplacien $Q(W)$ se définit par :

$$\mathcal{Q}_{uv}(W) = \begin{cases} -a_{uv} & 1 \leq u < v \leq n \\ \sum_{w=1}^{n} a_{wv} & 1 \leq u = v \leq n. \end{cases}$$

Alors pour tout u et v :

$$|Q^*_{uv}(W)| = \sum_{T \in \mathcal{T}} \prod_{\{k,l\} \in E_T} a_{kl}$$

Calcul de la probabilité conditionnelle

Théorème du matrix tree (étendu aux réels, [?])

Pour toute matrice symétrique $W = (a_{kl})_{k,l}$, son Laplacien $Q(W)$ se définit par :

$$\mathcal{Q}_{uv}(W) = \begin{cases} -a_{uv} & 1 \leq u < v \leq n \\ \sum_{w=1}^{n} a_{wv} & 1 \leq u = v \leq n. \end{cases}$$

Alors pour tout u et v :

$$|Q_{uv}^{*}(W)| = \sum_{T \in \mathcal{T}} \prod_{\{k,l\} \in E_T} a_{kl}$$

$$\begin{aligned} \mathbb{P}((k,l) \in T|Y) &= \sum_{T \in \mathcal{T}:(k,l) \in T} \mathbb{P}(T|Y) = \frac{\sum_{(k,l) \in T} \mathbb{P}(T)\mathbb{P}(Y|T)}{\sum_{T} \mathbb{P}(T)\mathbb{P}(Y|T)} \\ &= \frac{\sum_{(k,l) \in T} \prod_{uv} \beta_{uv}\psi_{uv}(Y)}{\sum_{T} \prod_{uv} \beta_{uv}\psi_{uv}(Y)} \\ &= 1 - \frac{|Q_{uv}^{*}(B\Psi^{-kl})|}{|Q_{uv}^{*}(B\Psi)|} \end{aligned}$$

Algorithme EM : étape M

But : optimiser les poids β_{kl}.

$$\underset{\beta_{kl}}{\arg\max} \left\{ \sum_{k,l \in V} \tau_{kl}(\log(\beta_{kl}) + \log(\psi_{kl})) - \log(B) - n \times cst \right\}$$

$$\frac{\partial \mathbb{E}_\theta[\log(\mathbb{P}(Y,T))|Y]}{\partial \beta_{kl}} = \frac{1}{\beta_{kl}}\tau_{kl} - \frac{1}{B}\frac{\partial B}{\partial \beta_{kl}}$$

Résultat de Meila [?]

En inversant un mineur du Laplacien Q, on définit la matrice symétrique M :

$$\begin{cases} M_{uv} = [\mathcal{Q}^{*-1}]_{uu} + [\mathcal{Q}^{*-1}]_{vv} - 2[\mathcal{Q}^{*-1}]_{uv} & u, v < n \\ M_{nv} = M_{vn} = [\mathcal{Q}^{*-1}]_{vv} & v < n \\ M_{vv} = 0. \end{cases}$$

On peut montrer que :

$$\frac{\partial |Q^*_{uv}(W)|}{\partial \beta_{kl}} = M_{kl} \times |Q^*_{uv}(W)|$$

Mise à jour des β_{kl}

$$\frac{\partial \mathbb{E}_\theta[\log(\mathbb{P}(Y,T))|Y]}{\partial \beta_{kl}} = \frac{1}{\beta_{kl}}\tau_{kl} - \frac{1}{B}\frac{\partial B}{\partial \beta_{kl}}$$

On rappelle que

$$B = \sum_{T\in\mathcal{T}} \prod_{k,l\in T} \beta_{kl}.$$

En utilisant le résultat de Meila pour dériver B, on obtient la formule de mise à jour à l'itération $h+1$:

$$\hat{\beta}_{kl}^{h+1} = \frac{\tau_{kl}^h}{M_{kl}^h}$$

Avec des données de comptage

Données :$(Y_{ij})_{i\in\{1,\ldots,n\},j\in\{1,\ldots,p+q\}}$: i échantillons de p variables observées, on suppose q variables supplémentaires non observées.

Modèle :

La loi Poisson log-Normale

$$\left.\begin{array}{ll} Z_i \; iid & \sim \mathcal{N}_{p+q}(\mu, \Sigma) \\ & (Y_{ij})_j \perp\!\!\!\perp |Z_i \\ Y_{ij}|Z_{ij} & \sim \mathcal{P}(e^{Z_{ij}}) \end{array}\right\} Y \sim \mathcal{PLN}(\mu, \Sigma)$$

Méthode :

- Inclure GGM dans le modèle Poisson-log normal
- Inférence variationnelle car $p(Z|Y)$ n'est pas calculable
- Prendre en compte un acteur manquant

Autres développements envisagés :

- Prise en compte de covariables
- Adapter le modèle à des données recueillies au cours du temps

Applications

- Comprendre les interactions (compétition, ...) entre les organismes
- Contrôle d'une espèce (d'un pathogène par exemple)

Collaborations

- Des projets d'écologie microbienne
  - INRA de Bordeaux avec C. Vacher (pathogène de la vigne et du chêne)
  - INRA de Rennes avec C. Mougel (rhizosphère)
- Projet ANR Hydrogen (analyse des données du projet TARA Océan)