Les modèles EDD : Une famille de modèles nuls génériques pour les réseaux écologiques

Tâm Le Minh, Sophie Donnet, François Massol, Stéphane Robin

MIA-Paris, INRAE

22 mars 2022

Statistiques au sommet de Rochebrune

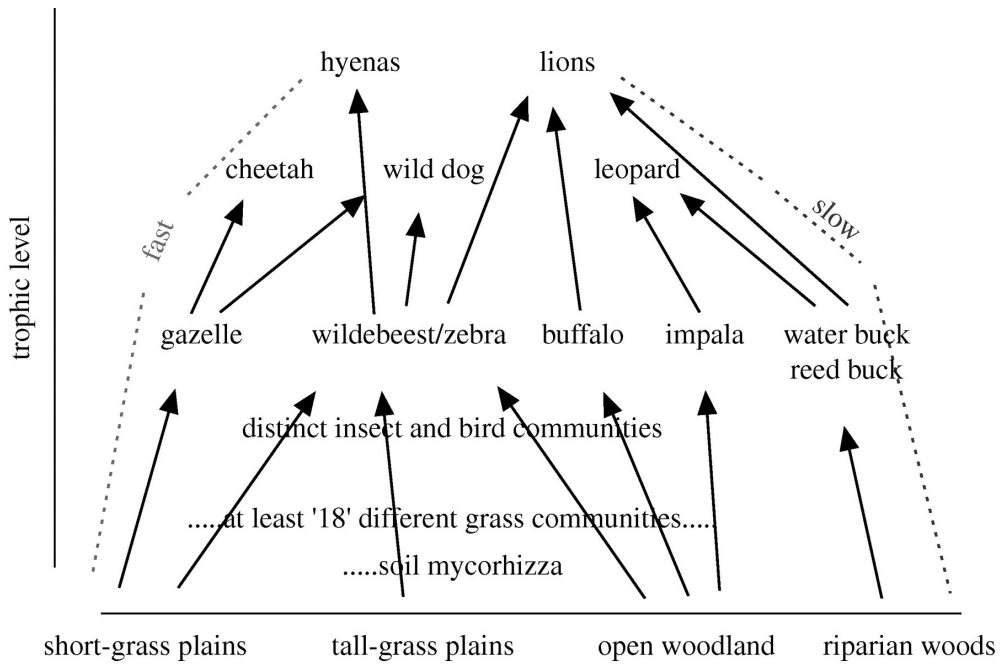
Réseaux d'interactions écologiques

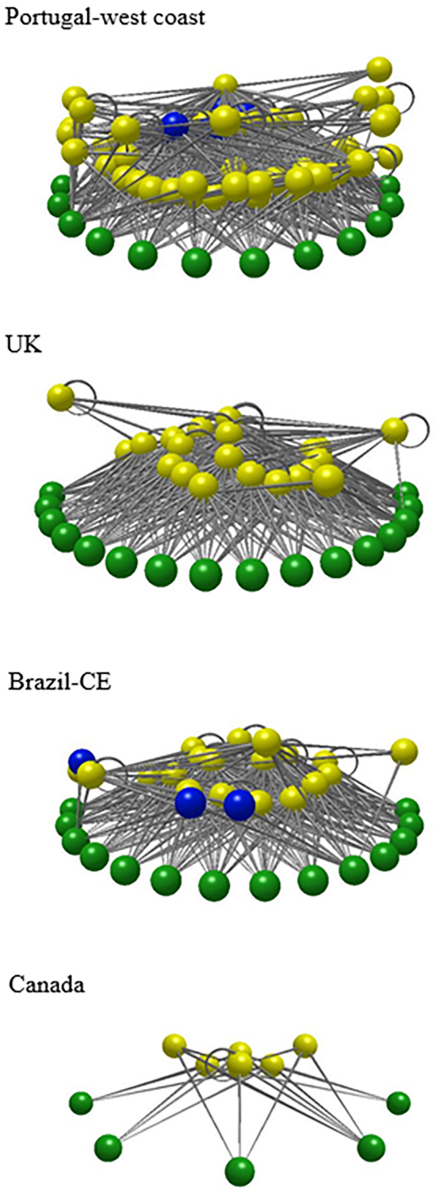
Interactions entre les espèces

$\rightarrow$ fonctionnement de l'écosystème

Variabilité des réseaux

$\rightarrow$ réponse aux perturbations extérieures

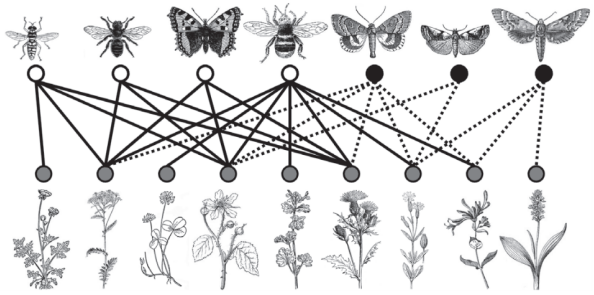


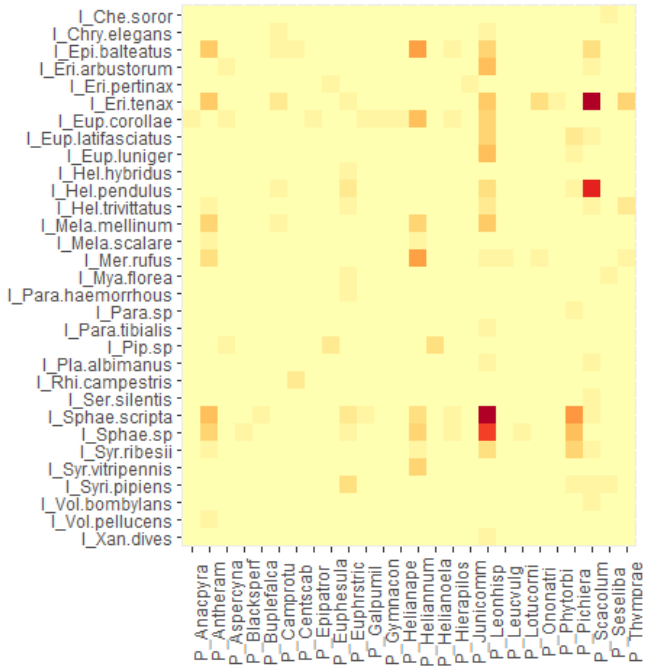


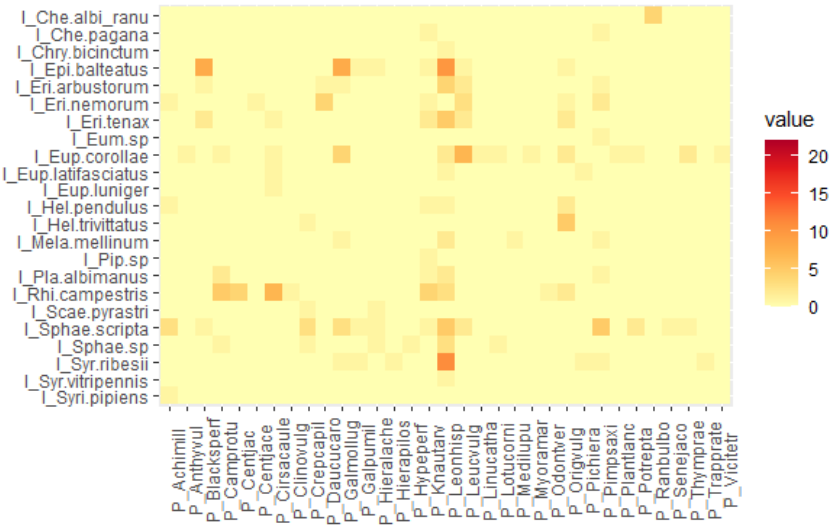
Matrice d'adjacence

Matrice d'adjacence Y :

Y_{ij} = interaction entre les espèces i et j





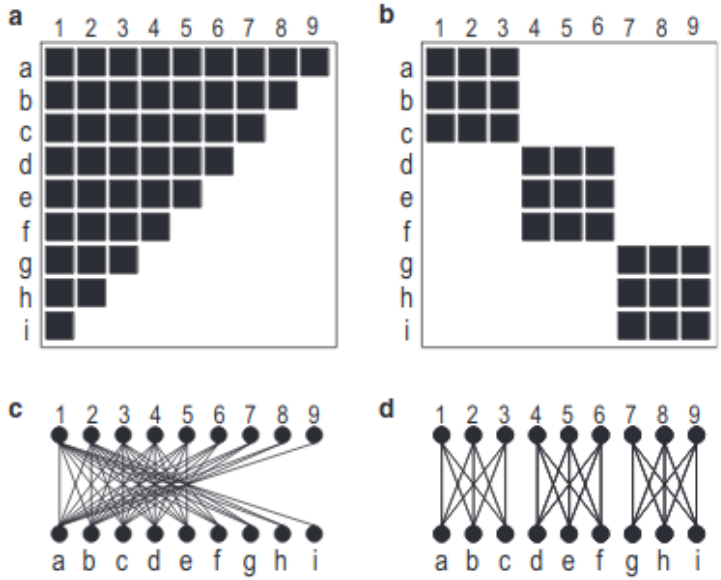


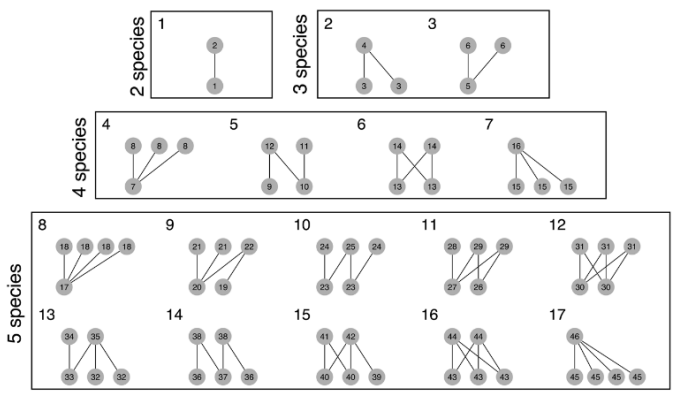
Représentation aussi valable pour les tables de présence-absence et d'abondance

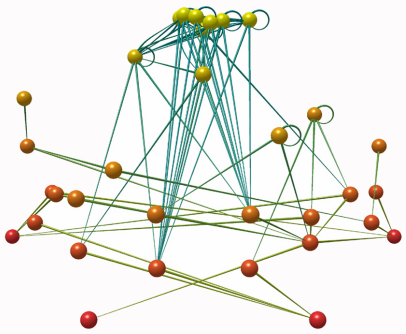
Étude des réseaux

Calcul de métriques globales :

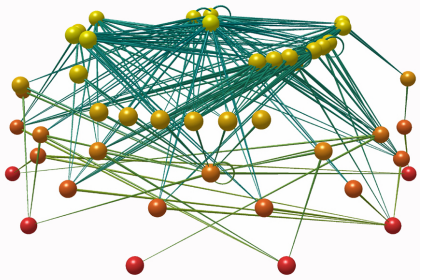
- Connectance, Nestedness, Modularité
- Indices de diversité
- Fréquences de motifs
- etc.







$S = 33$, $L = 99$, $C = 0.091$
$TL = 2.84$, Max$TL = 4.36$



$S = 48$, $L = 249$, $C = 0.108$
$TL = 2.72$, Max$TL = 3.78$

Utilisation de modèles nuls

Test d'hypothèse avec une statistique :

- On décide d'un niveau de significativité α,
- On calcule la distribution $\mathcal{F}_0$ de la statistique associée à l'hypothèse nulle $\mathcal{H}_0$,
- On construit une zone de rejet en fonction de $\mathcal{F}_0$ et de α,
- Si la statistique observée est dans la zone de rejet, alors l'hypothèse $\mathcal{H}_0$ est rejetée.

Le modèle nul génère des réseaux aléatoires utilisés pour calculer la distribution nulle $\mathcal{F}_0$ de la statistique.
$\rightsquigarrow$ $\mathcal{H}_0$ est donc déterminée par les hypothèses du modèle nul.

Modèles de configuration (à degrés fixés)

Connor et Simberloff (1979) : conservent les degrés/sommes des lignes et des colonnes

Exemple : algorithmes de swapping

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \Longleftrightarrow \quad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Interprétation écologique : l'hétérogénéité de spécialisation entre espèces crée le patron d'intérêt.

Cependant,

- Quelle distribution des réseaux générés ?
- Pourquoi conserver exactement les degrés ?

Modèles à degrés attendus fixés

Contraintes sur les espérances des degrés : modèles probabilistes

- Vazquez et Aizen (2003) : probabilité d'interaction **proportionnelle** à la fréquence des interactions de la ligne et de la colonne (modèle produit)
  $\rightsquigarrow$ on ne dispose pas de la fréquence des interactions + modèle très complexe
- Bascompte et al. (2003) : probabilité d'interaction égale à la moyenne des **probabilités d'occupation (degrés)** de la ligne et de la colonne
  $\rightsquigarrow$ les espérances des degrés ne correspondent pas aux degrés observés + pas de proportionnalité
  $\rightsquigarrow$ on peut imaginer une version produit de ce modèle

$$\mathbb{P}(Y_{ij} = 1) = \lambda \times w_i^{(r)} \times w_j^{(c)}$$

Modèle à distribution de degrés attendus (EDD)

Réseau défini par les distributions des **degrés** (sommes des arêtes issues des nœuds) **attendus** des lignes et des colonnes

Notations

- Densité du réseau : $\lambda \in]0, 1[$
- "Degré attendu" d'une ligne : $f(U_i),\ U_i \sim \mathcal{U}[0,1],\ \int f = 1$
- "Degré attendu" d'une colonne : $g(V_j),\ V_j \sim \mathcal{U}[0,1],\ \int g = 1$

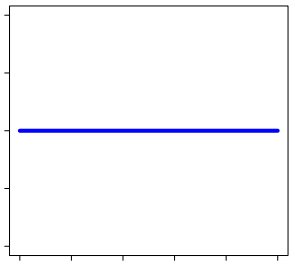
$$\begin{aligned} U_i, V_j &\overset{iid}{\sim} \mathcal{U}[0,1] \\ Y_{ij} \mid U_i, V_j &\sim \mathcal{B}(\lambda f(U_i) g(V_j)) \end{aligned}$$

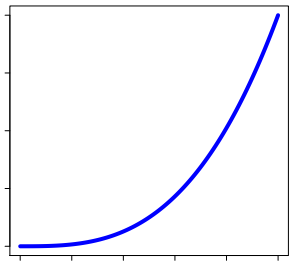
En tant que modèle nul, $\mathcal{H}_0$ est bien identifié :

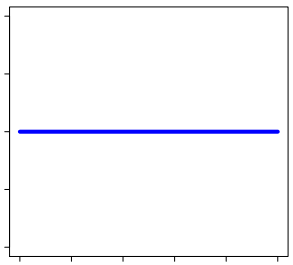
- les espèces tirent leurs degrés attendus dans des distributions,
- c'est un modèle produit.

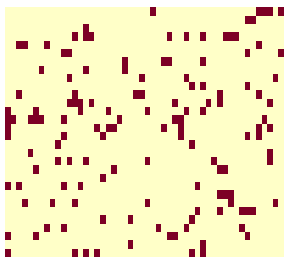
Modèle EDD

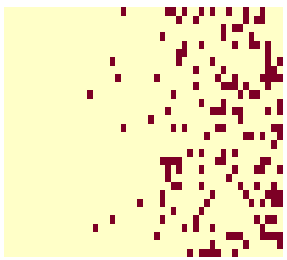
$$\begin{aligned} U_i, V_j &\overset{iid}{\sim} \mathcal{U}[0,1] \\ Y_{ij} \mid U_i, V_j &\sim \mathcal{B}(\lambda f(U_i) g(V_j)) \end{aligned}$$

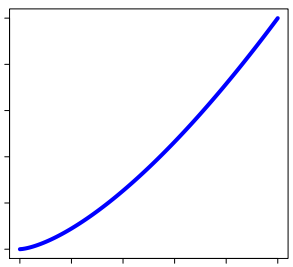
$g_0(v) =$ 

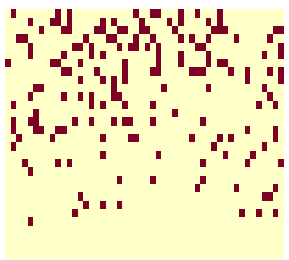
$g(v) =$ 

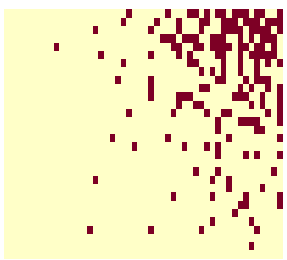
$f_0(u) =$ 





$f(u) =$ 





Une version pondérée du modèle EDD

Réseau défini par les distributions des **degrés** (sommes des arêtes issues des nœuds) **attendus** des lignes et des colonnes

Notations

- Densité du réseau : $\lambda \in \mathbb{R}_+^*$
- "Degré attendu" d'un insecte : $f(U_i),\ U_i \sim \mathcal{U}[0,1],\ \int f = 1$
- "Degré attendu" d'une plante : $g(V_j),\ V_j \sim \mathcal{U}[0,1],\ \int g = 1$

$$\begin{array}{rcl} U_i, V_j & \overset{iid}{\sim} & \mathcal{U}[0,1] \\ Y_{ij} \mid U_i, V_j & \sim & \mathcal{P}(\lambda f(U_i) g(V_j)) \end{array}$$

En tant que modèle nul, $\mathcal{H}_0$ est bien identifié :

- les espèces tirent leurs degrés attendus dans des distributions,
- c'est un modèle produit,
- les fréquences d'interactions suivent une loi de Poisson.

Propriété d'échangeabilité

Le modèle EDD est un modèle échangeable ligne-colonne : la loi jointe de la matrice est invariante par permutation des lignes ou des colonnes.

Pour des permutations σ_1 et σ_2 :

- $(Y_{1\bullet}, ..., Y_{m\bullet}) \overset{\mathcal{L}}{=} (Y_{\sigma_1(1)\bullet}, ..., Y_{\sigma_1(m)\bullet})$,
- $(Y_{\bullet 1}, ..., Y_{\bullet n}) \overset{\mathcal{L}}{=} (Y_{\bullet\sigma_2(1)}, ..., Y_{\bullet\sigma_2(n)})$,

où $Y_{i\bullet} := (Y_{i1}, ..., Y_{in})$, $Y_{\bullet j} := (Y_{1j}, ..., Y_{mj})$.

Propriété d'échangeabilité

En général, les statistiques étudiées donnent des informations sur la topologie globale du réseau (nestedness, modularité, fréquences de motifs). Elles ne dépendent pas du nom des espèces.

Toutes les espèces jouent le même rôle dans l'étude : on peut omettre la taxonomie, i.e. considérer que si on permute les lignes ou les colonnes de la matrice d'adjacence, elle représente toujours le même réseau.

Conséquences :

1. On peut utiliser des modèles échangeables ligne-colonne.
2. Certaines statistiques peuvent s'écrire sous la forme d'U-statistiques.

$\rightsquigarrow$ Le modèle EDD est adapté à ces études.

U-statistiques

Hoeffding (1948) : $(X_1, ..., X_n)$ variables i.i.d.

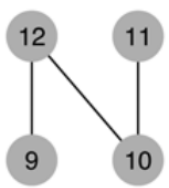
$$U = r! \binom{n}{r}^{-1} \sum_{1 \leq i_1 \neq ... \neq i_r \leq n} h(X_{i_1}, ..., X_{i_r}).$$

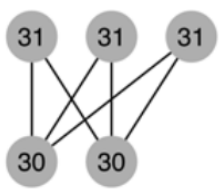
U-statistique sur une matrice $m \times n$

$$U = p!q! \left[\binom{m}{p} \binom{n}{q} \right]^{-1} \sum_{i_1 \neq ... \neq i_p}^{m} \sum_{j_1 \neq ... \neq j_q}^{n} h(Y_{\{i_1,...,i_p;j_1,...,j_q\}})$$

Exemple : fréquences de motifs

$$h(Y_{\{1,2;1,2\}}) = Y_{11} Y_{12} Y_{21} (1 - Y_{22})$$
$$h(Y_{\{1,2;1,2,3\}}) = Y_{11} Y_{12} Y_{13} Y_{21} Y_{22} Y_{23}$$





Estimation du modèle EDD

Rappel : Modèle EDD bipartite pondéré

$$\begin{aligned} U_i, V_j &\overset{iid}{\sim} \mathcal{U}[0,1] \\ Y_{ij} \mid U_i, V_j &\sim \mathcal{P}(\lambda f(U_i) g(V_j)) \end{aligned}$$

où :

- $\lambda = \mathbb{E}[Y_{ij}]$
- $\int f = \int g = 1$, $\int f^k = F_k$, $\int g^k = G_k$.

Quelques propriétés :

$\rightarrow \mathbb{E}[Y_{i_1 j_1}^2 - Y_{i_1 j_1}] = \lambda^2 F_2 G_2$

$\rightarrow \mathbb{E}[Y_{i_1 j_1} Y_{i_1 j_2}] = \lambda^2 F_2$

$\rightarrow \mathbb{E}[Y_{i_1 j_1} Y_{i_2 j_1}] = \lambda^2 G_2$

$\rightsquigarrow$ Fonctions sur le quadruplet : $h(Y_{\{i_1,i_2;j_1,j_2\}}) = h(Y_{i_1 j_1}, Y_{i_1 j_2}, Y_{i_2 j_1}, Y_{i_2 j_2})$

U-statistiques

U-statistique sur les quadruplets

$$U^h_{m,n} = \left[\binom{m}{2}\binom{n}{2}\right]^{-1} \sum_{i_1<i_2}^{m} \sum_{j_1<j_2}^{n} h(Y_{\{i_1,i_2;j_1,j_2\}})$$

Estimateur $\widehat{\theta}_N := U^h_{cN,(1-c)N} \rightsquigarrow \mathbb{E}[\widehat{\theta}_N] = \mathbb{E}[h(Y_{\{i_1,i_2;j_1,j_2\}})] = \theta$

TCL pour les modèles échangeables ligne-colonne

$$\sqrt{\frac{N}{V}}(\widehat{\theta}_N - \theta) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

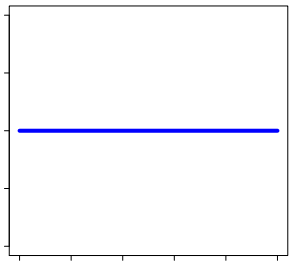
Les U-statistiques permettent de faire de l'inférence statistique avec un minimum d'hypothèses

- Estimation
- Intervalles de confiance
- Tests de comparaison

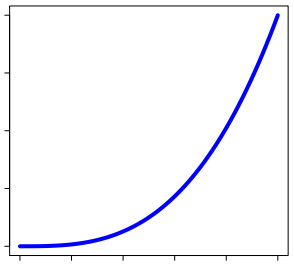
Exemple : test sur f

$\mathcal{H}_0 : f \equiv 1$ contre $\mathcal{H}_1 : f \not\equiv 1$
$(F_2 = 1$ contre $F_2 > 1)$

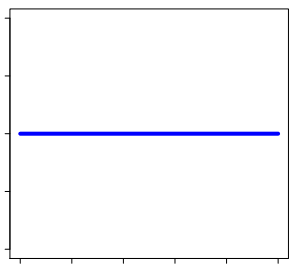
$g_0(v) =$

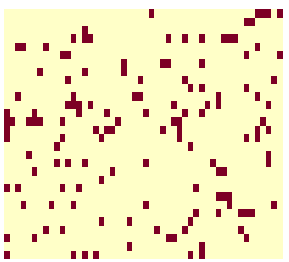


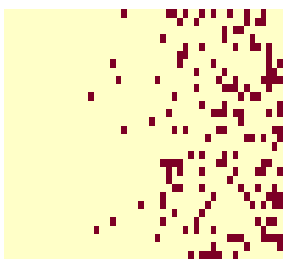
$g(v) =$



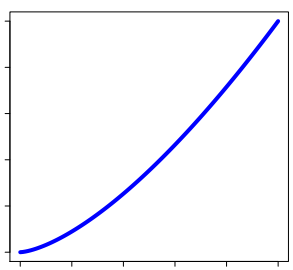
$f_0(u) =$

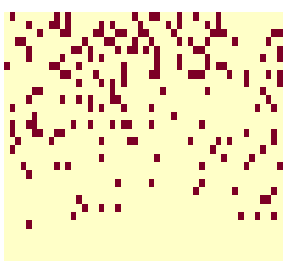


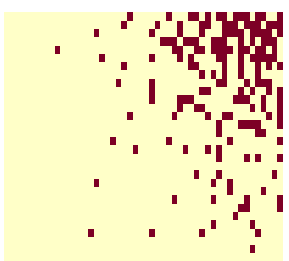




$f(u) =$







Exemple : test sur f

Étape 1 : Choix des noyaux des U-statistiques

- $h_1(Y_{\{i_1,i_2;j_1,j_2\}}) = \frac{1}{2}(Y_{i_1j_1}Y_{i_1j_2} + Y_{i_2j_1}Y_{i_2j_2})$

  $\mathbb{E}\left[h_1(Y_{\{i_1,i_2;j_1,j_2\}})\right] = \lambda^2 F_2$
- $h_2(Y_{\{i_1,i_2;j_1,j_2\}}) = \frac{1}{2}(Y_{i_1j_1}Y_{i_2j_2} + Y_{i_1j_2}Y_{i_2j_1})$

  $\mathbb{E}\left[h_2(Y_{\{i_1,i_2;j_1,j_2\}})\right] = \lambda^2$

Étape 2 : Propriétés de l'estimateur

Normalité asymptotique + théorème de Slutsky

$$\sqrt{\frac{N}{V^{h_1}}}\left(U_N^{h_1} - U_N^{h_2}\right) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

Estimateur de V^{h_1} ?

Exemple : test sur f

Dans le modèle EDD et pour h_1, sous $\mathcal{H}_0$,

$$V^{h_1} = \frac{4\lambda^4}{1-c}(G_2 - 1)$$

Pour estimer V^{h_1}, il faut encore estimer G_2 :

- $h_3(Y_{\{i_1,i_2;j_1,j_2\}}) = \frac{1}{2}(Y_{i_1j_1}Y_{i_2j_1} + Y_{i_1j_2}Y_{i_2j_2})$

  $\mathbb{E}[h_3(Y_{\{i_1,i_2;j_1,j_2\}})] = \lambda^2 G_2$

Étape 3 : Estimateur consistant de la variance

$$\widehat{V}_N^{h_1} = \frac{4}{1-c}(U_N^{h_2})^2 \left[\frac{U_N^{h_3}}{U_N^{h_2}} - 1\right]$$

Exemple : test sur f

Conclusion

Encore une application de Slutsky,

$$\sqrt{\frac{N}{\widehat{V}_N^{h_1}}}\left(U_N^{h_1} - U_N^{h_2}\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

$\mathcal{H}_0 : f \equiv 1$ contre $\mathcal{H}_1 : f \not\equiv 1 \Rightarrow$ On rejette $\mathcal{H}_0$ au niveau α si la statistique de test

$$T_N := \sqrt{\frac{N}{\widehat{V}_N^{h_1}}}\left(U_N^{h_1} - U_N^{h_2}\right) > q_{1-\alpha}.$$

Bilan

Les modèles EDD sont :

- des modèles nuls probabilistes génératifs
  $\rightsquigarrow \mathcal{H}_0$ correspond à une hypothèse écologique,

- des modèles échangeables ligne-colonne
  $\rightsquigarrow$ résultats de convergence des U-statistiques,

- des modèles semi-paramétriques mais les U-statistiques ne nécessitent pas de connaître les distributions de degrés
  $\rightsquigarrow$ possibilité de faire de l'inférence avec le minimum d'hypothèses.

Bibliographie


Bascompte, J., Jordano, P., Melián, C. J., & Olesen, J. M. (2003).
The nested assembly of plant–animal mutualistic networks. *Proceedings of the National Academy of Sciences*.


Connor, E. F., & Simberloff, D. (1979).
The assembly of species communities : chance or competition ? *Ecology*.


Hoeffding, W. (1948).
A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*.


Vázquez, D. P., & Aizen, M. A. (2003).
Null model analyses of specialization in plant–pollinator interactions. *Ecology*.

Merci de votre attention !