Inférence de réseaux écologiques par réseau bayésien dynamique

Étienne Auclair

INRA - Unité MIAT

July 3, 2015

Formation

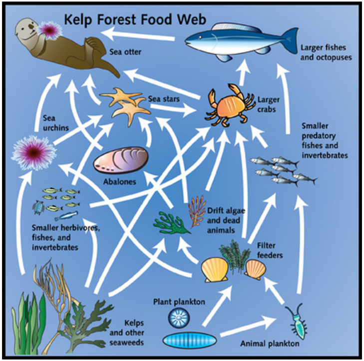
Formation

- DUT **St**atistique et **I**nformatique **D**écisionnelle - Carcassonne
- L3 Économie et Informatique - TSE
- Université Lyon 2 Lumière - Département Informatique et Statistique
  - Master **E**xtraction des **C**onnaissances à partir des **D**onnées

Contexte biologique et objectif

Contexte

- Gestion de la biodiversité dans un réseau écologique
- Difficulté : les interactions sont mal connues
- Influence de la protection



Objectif

Apprendre la structure d'un réseau écologique à partir de données de présence/absence temporelles

Données

Données réelles disponibles

- Fournies par Jennifer Caselle et Laura Dee (Univ. of California, Santa Barbara)
- Observation d'espèces marines dans des forêts de kelp au large de la Californie
- 164 espèces d'un réseau connu observées pendant 14 ans dans 4 zones différentes

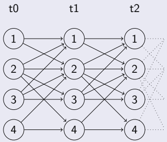
Réseaux bayésiens

Réseau bayésien

- Dépendances entre variables : Graphe orienté sans circuit (DAG)
- Description quantitative des dépendances : Tables de probabilité conditionnelles (Paramètres du modèle)

Réseau bayésien dynamiques

- Mesure un phénomène récurrent (ex : temporel)
- Extension d'une chaîne de markov : l'état des espèces à un temps t ne dépend que de son état à $t - 1$
- Processus stationnaire : les probabilités de transitions ne changent pas au cours du temps



Apprentissage de structure de réseau bayésien

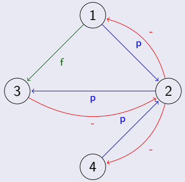
Apprentissage par score

- Calcul de score à partir des paramètres du modèle (ex : BIC)
- Algorithme glouton
  - Étape 1 : Estimation des paramètres à graphe G connu
  - Étape 2 : Recherche d'un graphe améliorant G selon le critère BIC
  - Retour à l'étape 1 jusqu'à convergence

Modèle

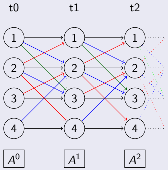
Modélisation de réseau écologique

- Graphe orienté
- Arcs labellisés selon l'interaction :
    - p : proie
    - f : facilitateur
    - - : influence négative



Modélisation des dynamiques des espèces

- Modélisation par réseau bayésien dynamique



Spécificité du stage

Modèle paramétré

- Pas de table de probabilité conditionnelles
  - Probabilité de survie différente de la probabilité de recolonisation
  - Un paramètre par type de relation
  - Probabilité différente selon l'état de protection

➢ Que deviennent les formules du maximum de vraisemblance ?

Arcs labellisés

➢ Comment apprendre les labels des arcs ?

Notations

Données

- $X_t^i \in \{1, 0\}$ présence ou absence de l'espèce i ($i \in \{1, ..., n\}$) au temps t ($t \in \{1, ..., T\}$).
- $A^t \in \{1, 0\}$ protection ou absence de protection de la zone d'observation au temps t.
- $\overset{int}{N_i^t}$ nombre de parents de type *int* de l'espèce i présents au temps t.

Paramètres

- Probabilité de recolonisation ε.
- Probabilités de réussite de chaque type d'influence ρ^p, ρ^f, ρ^-.
- Paramètre pénalisant l'absence de protection d'une zone : μ.

Probabilités

Recolonisation

Espèce absente au temps $t-1$: probabilité de recolonisation au temps t

- $P(X_i^t = 1 | X_i^{t-1} = 0, A^{t-1} = 1) = \varepsilon$
- $P(X_i^t = 1 | X_i^{t-1} = 0, A^{t-1} = 0) = \mu\varepsilon$

Survie

Espèce présente au temps $t-1$: probabilité de survie au temps t

- $P(X_i^t = 1 | X_i^{t-1} = 1, A^{t-1} = 1) =$
$$\left(1 - (1-\rho^p)^{\overset{p}{N}_i^t}\right)\left(1 - (1-\rho^f)^{\overset{f}{N}_i^t}\right)(1-\rho^-)^{\overset{-}{N}_i^t}$$
- $P(X_i^t = 1 | X_i^{t-1} = 1, A^{t-1} = 0) =$
$$\mu\left(1 - (1-\rho^p)^{\overset{p}{N}_i^t}\right)\left(1 - (1-\rho^f)^{\overset{f}{N}_i^t}\right)(1-\rho^-)^{\overset{-}{N}_i^t}$$

Déroulement du stage

Estimation des paramètres

- *Écriture de la vraisemblance et de son gradient*
- *Implémentation des formules de la vraisemblance et du gradient*
- **Recherche du maximum de vraisemblance**

Apprentissage de la structure

- Amélioration de la structure du graphe
- Algorithme itératif : alterner estimation et amélioration

Validation

- *Implémentation d'un simulateur du modèle*
- **Validation sur données simulées**
- Validation sur données réelles

Gradient

$$\frac{\partial}{\partial \varepsilon} = \left(\#_{0\to 1}^{a=1} + \#_{0\to 1}^{a=0}\right)\frac{1}{\varepsilon} - \left(\#_{0\to 0}^{a=1}\right)\frac{1}{1-\varepsilon} - \left(\#_{0\to 0}^{a=0}\right)\frac{\mu}{1-\mu\varepsilon}$$

$$\frac{\partial}{\partial \rho^p} = \sum_{\substack{m,n,p \\ tq\ 0<m+n+p\leq K}} \frac{mx^{m-1}\left(\alpha(m,n,p)+\beta(m,n,p)\right)}{1-x^m} - \gamma(m,n,p)\frac{mx^{m-1}\left(1-y^n\right)z^p}{1-\left(1-x^m\right)\left(1-y^n\right)z^p} - \delta(m,n,p)\frac{\mu mx^{m-1}\left(1-y^n\right)z^p}{1-\mu\left(1-x^m\right)\left(1-y^n\right)z^p}$$

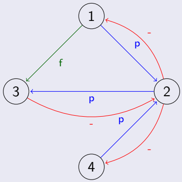
$$\frac{\partial}{\partial \rho^f} = \sum_{\substack{m,n,p \\ tq\ 0<m+n+p\leq K}} \frac{ny^{n-1}\left(\alpha(m,n,p)+\beta(m,n,p)\right)}{1-y^n} - \gamma(m,n,p)\frac{\left(1-x^m\right)ny^{n-1}z^p}{1-\left(1-x^m\right)\left(1-y^n\right)z^p} - \delta(m,n,p)\frac{\mu\left(1-x^m\right)ny^{n-1}z^p}{1-\mu\left(1-x^m\right)\left(1-y^n\right)z^p}$$

$$\frac{\partial}{\partial \rho^p} = \sum_{\substack{m,n,p \\ tq\ 0<m+n+p\leq K}} -\frac{p\left(\alpha(m,n,p)+\beta(m,n,p)\right)}{z} + \gamma(m,n,p)\frac{\left(1-x^m\right)\left(1-y^n\right)pz^{p-1}}{1-\left(1-x^m\right)\left(1-y^n\right)z^p} + \delta(m,n,p)\frac{\mu\left(1-x^m\right)\left(1-y^n\right)pz^{p-1}}{1-\mu\left(1-x^m\right)\left(1-y^n\right)z^p}$$

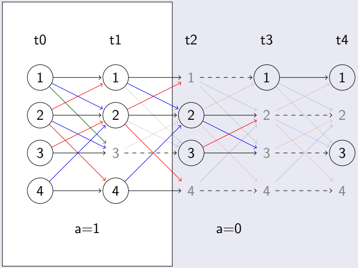
$$\frac{\partial}{\partial \mu} = \left(\#_{0\to 1}^{a=0} + \sum_{\substack{m,n,p \\ tq\ 0<m+n+p\leq K}} \beta(m,n,p)\right)\frac{1}{\mu} - \left(\#_{0\to 0}^{a=0}\right)\frac{\varepsilon}{1-\mu\varepsilon} - \sum_{\substack{m,n,p \\ tq\ 0<m+n+p\leq K}} \delta(m,n,p)\frac{\left(1-x^m\right)\left(1-y^n\right)z^p}{1-\mu\left(1-x^m\right)\left(1-y^n\right)z^p}$$

Simulation

Graphe



Données simulées



Conclusion

Pour les écologues

- Amélioration de la connaissance des interactions biologiques
- Mesure de l'impact de la protection

Poursuite

- Gestion de la biodiversité lorsque les interactions sont mal connues

➢ Comment combiner gestion et apprentissage ?