Learning ecological network structure using parametrized Dynamic Bayesian Network

Étienne Auclair

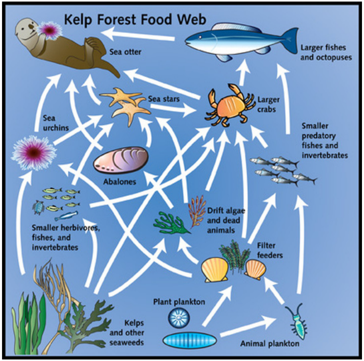
INRA - Unité MIAT

October 12, 2016

Ecological context and objective

Context

- Management of biodiversity within an ecological network
- Interactions are poorly known
- Protection of certain areas



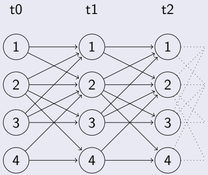
Objective

Developing a method for learning the structure of an ecological network using presence/absence temporal data

Probabilistic network learning

Bayesian network

- Bayesian network
  - Directed acyclic graph
  - Conditional probability tables
- Dynamic Bayesian network (DBN)
  - Recurrent phenomenon (temporal...)
  - Stationary Markov process



Learning the structure of BN

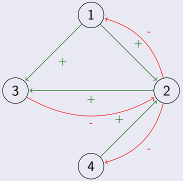
Score learning methods

- Score calculated using the parameters of the model (BIC, BDe)
- Greedy algorithm
  - Step 1 : Estimating the parameters with a known graph G
  - Step 2 : Search of a new graph improving score
  - Back to step 1 until convergence

DBN model of an ecological network

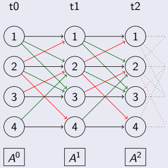
Ecological network

- Directed graph
- Edges labelled according to the type of interaction :
  - $+$: Positive influence
  - - : Negative influence



Modelling the dynamic of the species

- Dynamic Bayesian Network model



Notations

Data

- $X_t^i \in \{1,0\}$ presence or absence of the species i ($i \in \{1, ..., n\}$) at time step t ($t \in \{1, ..., T\}$).
- $A^t \in \{1,0\}$ protection or absence of protection at time step t.
- $N_{i,l}^t$ number of "l" labelled parents of the species i present at time step t.

Parameters

- Recolonization probability ε.
- Probability of success of each influence ρ^+, ρ^-.
- Penalization for unprotected moments : μ.

Probabilities

Recolonization

Species absent at moment $t-1$: probability of recolonization at time step t

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

Survival

Species present at moment $t-1$: probability of survival at time step t

- $P(X_i^t = 1 | X_i^{t-1} = 1, A^{t-1} = 1) = \left(1 - (1-\rho^+)^{N_{i,+}^t}\right)(1-\rho^-)^{N_{i,-}^t}$
- $P(X_i^t = 1 | X_i^{t-1} = 1, A^{t-1} = 0) = \mu\left(1 - (1-\rho^+)^{N_{i,+}^t}\right)(1-\rho^-)^{N_{i,-}^t}$

Expression of the likelihood

$\log P_{\mathcal{LG}_{\rightarrow},\theta}(x^2, \ldots, x^T \mid x^1, a) = \sum_{i=1}^{n} score(i)$

Learning a Parametrized labelled DBN

Parametrized labelled DBN

- No conditional probability tables
  - Independent recolonization probabilities
  - A parameter per interaction type
  - Decreased probability when there is no protection
- No explicit expression of the maximum likelihood
- How to learn labelled edges ?

Learning P-DBN by score-based method

- Fixed number of parameters : likelihood as score
- Greedy algorithm
  - Step 1 : Parameters estimation by likelihood maximization
  - Step 2 : Graph structure learning by 0-1 integer linear programming
  - Back to step 1 until convergence

Optimal graph structure

Integer linear programming (ILP) 0-1

- Linearisation of the problem : addition of binary variables defined by linear constraints
- Optimization of the score using ILP
- One independent ILP per species

Characteristics of the ILP

For n species, T time steps and k parents at most :

- Number of variables : $\left(3 \cdot n + 1 + T \cdot \left(\frac{k^2}{2} + \frac{3 \cdot k}{2} + 8\right)\right)$ for each species.
- Number of constraints : $\left(n + 1 + T \cdot \left(2 \cdot k^2 + 6 \cdot k + 21\right)\right)$ for each species.

Simulated data

Network and covariates

- Extract from real network : subgraph where no species have more than k parents
- Observed on $T = 30$ years
- The last 18 years are *protected*

Parameters

Every set of parameters configuration for the values $\{0.2, 0.8\}$

$$1: \quad \{\varepsilon = 0.2, \rho^{+} = 0.2, \rho^{-} = 0.2, \mu = 0.2\}$$

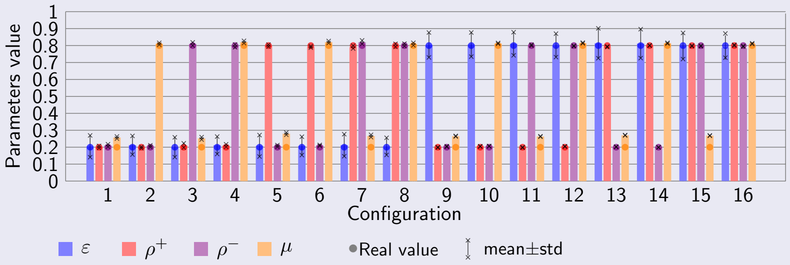
$$\dots$$

$$16: \quad \{\varepsilon = 0.8, \rho^{+} = 0.8, \rho^{-} = 0.8, \mu = 0.8\}$$

Estimation of the parameters

Network : $k = 2$, $n = 18$. 150 simulations.

Figure : Quality of the parameters estimation step

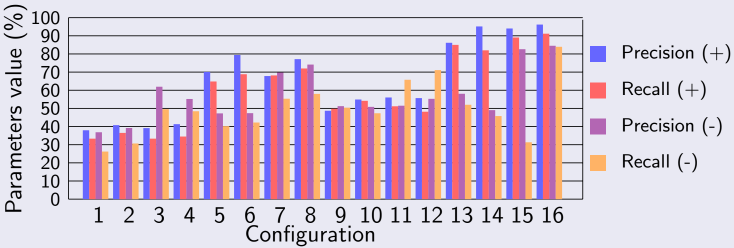


- Estimated parameters close to real parameters
- Better estimation for higher parameter value

Learning the structure

Network : $k = 2$, $n = 4$. 150 simulations.

Figure : Quality of the structure learning step



- Precision= $\frac{\text{edges correctly learnt}}{\text{edges learnt}}$
- Recall= $\frac{\text{edges correctly learnt}}{\text{edges present in the \textit{original graph}}}$

P-DBN learning algorithm

Network : $k = 4$, $n = 45$. 40 simulations.

Global results

- Average precision : $14.07\%(+); 17.96\%(-)$.
- Average recall : $29.53\%(+); 19.09\%(-)$.

- Learning on one presence/absence data is not efficient
- Does our method fail to learn the interactions ?

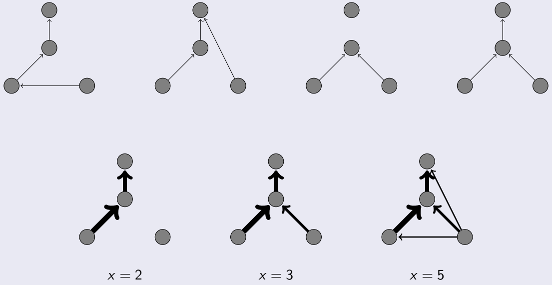
Modal graph

Definition

Consensus over every simulations of the x most learnt edges

Learnt graph

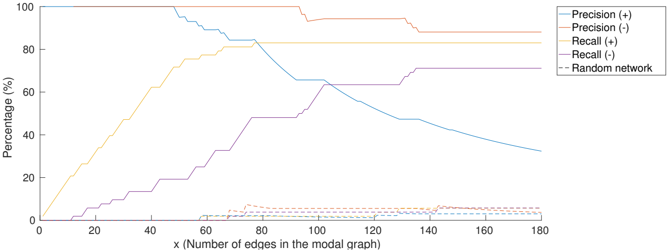
Modal graph of x edges



Modal graph results

Modal graph of the x most often learnt edges amongst 40 simulations

Figure : Performances of the modal graph given x



How to apply this method on real data ?

Kelp forest dataset

PISCO survey

- Abundance of fishes, macroalgae and invertebrates
- 4 sites of observation with different status of protection
- 15 years of monitoring (2000-2014)
- 250 species monitored
- Some interactions are known

Abundance to presence/absence

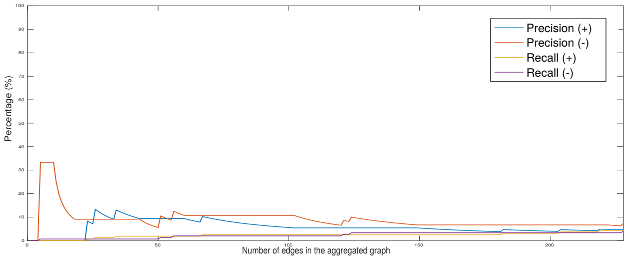
- Building several presence/absence dataset ?
- Thresholds on scaled abundance data

Structure learning results on real data

Data used

- Selection of $n = 38$ species with known interactions
- Area protected since 2003 (15 years of observation - 3 unprotected 12 protected)

Figure : Performance on the modal graph for real data

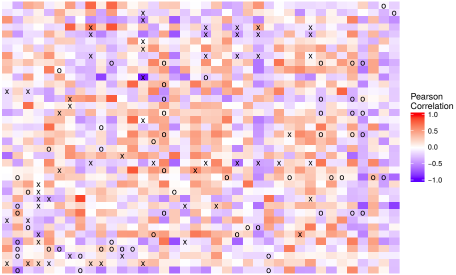


Analysis on the results

Why those results ?

- Did we miss some key interacting species ?
- Is the dynamic of the species influenced by the interactions ?

Figure : Heatmap of the coefficient of correlation between the time series of the species. o : Positive influence - x : Negative influence



Conclusion

Parameterized Dynamic bayesian network

- DBN with a given set of parameters
- Structure learning using ILP

Results

- Learning one one dataset is hard
- Difficulties to learn the structure on real data

Perspectives

- Management of the biodiversity within an unknown ecological network
- Managing while learning