Inferring higher-order co-occurrence patterns and **simplicial complexes** from presence/absence data

Xavier Roy-Pomerleau[1, 2], Louis J. Dubé[1, 2], Patrick Desrosiers[1, 2, 3]

1. Département de physique, génie physique et d'optique, Université Laval, Québec, Canada
2. Centre Interdisciplinaire en Modélisation Mathématique de l'Université Laval, Québec, Canada
3. Centre de recherche CERVO, Québec, Canada

xavier.roy-pomerleau.1@ulaval.ca

https://github.com/pdesrosiers/HOLMES

How to infer higher-order co-occurrence patterns and simplicial complexes from presence/absence data?

By using log-linear models and hypothesis testing!

	Site 1	Site 2	Site 3	Site 4	Site 5
Species A	1	0	1	1	0
Species B	1	1	0	1	0
Species C	0	0	0	0	1

Step 1 : Fill a contingency table for each pair

Contingency table : Table in which we count how many times a specific presence/absence situation appeared in the data.

	Site 1	Site 2	Site 3	Site 4	Site 5
Species A	1	0	1	1	0
Species B	1	1	0	1	0

	Species $B=0$	Species $B=1$	Total
Species A $=0$	$x_{00}=1$	$x_{01}=1$	$x_{0+}=2$
Species A $=1$	$x_{10}=1$	$x_{11}=2$	$x_{1+}=3$
Total	$x_{+0}=2$	$x_{+1}=3$	$N=5$

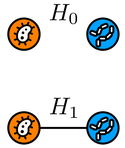
Step 2 : Set hypotheses and corresponding log-linear models

H_0: Species i and j occur independently.

$$\log(m_{ij}) = u + u_i^A + u_j^B$$

H_1: Species i and j are correlated.

$$\log(m_{ij}) = u + u_i^A + u_j^B + u_{ij}^{AB}$$



Contingency tables are instances of a **multinomial distribution**. The log-likelihood of such distribution is given by

$$\log\left(\frac{N!}{\prod_{ij} x_{ij}!}\right) + \sum_{i,j} x_{ij}\log(m_{ij}) - N\log(N),$$

where : N is the total number of observations;
x_{ij} are the cell entries in the contingency table;
m_{ij} are the expected counts in the multinomial distribution.

Step 3 : Find expected values under H_0

We rewrite the log-likelihood of the sampling distribution as

$$\log\left(\frac{N!}{\prod_{ij} x_{ij}!}\right) + \sum_{ij} x_{ij}\left(u + u_i^A + u_j^B\right) - N\log(N),$$

and design an iterative procedure to find the maximum likelihood estimates.

	Species $B=0$	Species $B=1$
Species A $=0$	$\hat{m}_{00}$	$\hat{m}_{01}$
Species A $=1$	$\hat{m}_{10}$	$\hat{m}_{11}$

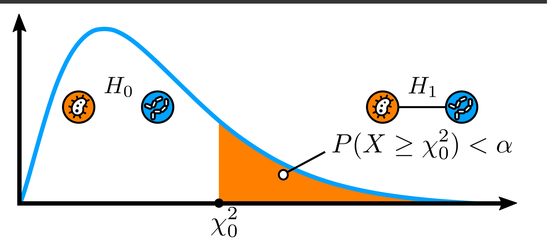
Where $\hat{m}_{ij}$ are the maximum likelihood estimates under H_0.

Step 4 : Test H_0 using χ^2 statistics

Using the χ_0^2 statistics, we measure how close our observations are from the expected values under H_0. We compute the statistics with

$$\chi_0^2 = \sum_{i,j} \frac{(x_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}.$$

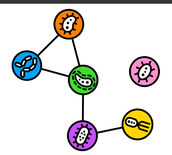
We reject the hypothesis with a significance level α if the probability of drawing χ_0^2 from a χ^2 distribution is smaller than α.



Step 5 : Repeat for each pair

By repeating for each pair, we infer a **network** of statistically significant co-occurrences!

Nodes : Observed species
Links : Probabilistic dependencies in the occurrence



When the **number of observations is low**, the statistics is not distributed as a χ^2 distribution and step 4 **will not give an accurate result**.

In that case we **need to** generate the **exact distribution** of the statistics for each pair.

Step 6 : Repeat for each triple with higher-order log-linear models

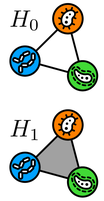
The **only extra steps** are to find the **new log-likelihood** and set the a**ppropriate hypotheses**.

H_0: Species i, j and k are dependent through pairwise dependencies.

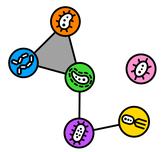
$$\log(m_{ijk}) = u + u_i^A + u_j^B + u_k^C + u_{ij}^{AB} + u_{ik}^{AC} + u_{jk}^{BC}$$

H_1: Species i, j and k form a higher-order co-occurrence pattern.

$$\log(m_{ijk}) = u + u_i^A + u_j^B + u_k^C + u_{ij}^{AB} + u_{ik}^{AC} + u_{jk}^{BC} + u_{ijk}^{ABC}$$



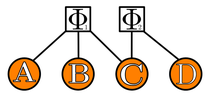
We obtain a **simplicial complex** with **higher-order co-occurrence patterns!**



Repeat this procedure for the highest possible order!

Validation of the inference method with a generative model

A **factor graph** is a bipartite graph that encodes the relationships between random variables via factor nodes. The probability of drawing a particular state for a set of random variables linked to the factor node is determined by the factor [2]. With $A, B, C \in \{0, 1\}$,



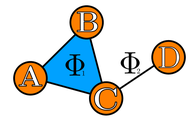
$$P(A,B,C) = \frac{\Phi_1(A,B,C)}{Z} = \frac{e^{-H(A,B,C)}}{Z},$$

where $H(A,B,C)$ is an energy function and Z is the partition function. We design each factor such that its logarithm can be mapped to a log-linear model. For the previous factor graph, we could choose

$$H(A,B,C) = \omega_1 ABC + \omega_2 AB(1-C) + ... + \omega_{n-1} B + \omega_n C\,,$$

where $\omega_1, ..., \omega_n$ are real numbers.

If designed carefully, a factor graph can be mapped to a simplicial complex.



Using a **rejection sampling** scheme and the total distribution of the factor graph, one can generate **synthetic observations**.

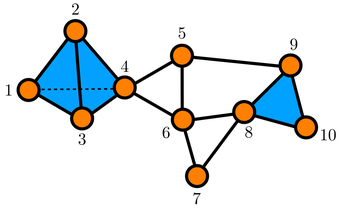
	Instance 1	Instance 2	Instance 3	Instance 4	Instance 5	...
A	1	0	1	1	0	...
B	1	1	0	1	0	...
C	0	0	0	0	1	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮

One can infer a simplicial complex from synthetic data and compare its structure to the original simplicial complex.

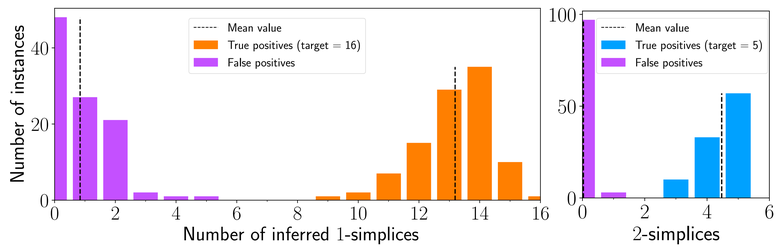
Results on synthetic and real datasets

Synthetic data

From the following simplicial complex representation of a factor graph, we generated 100 presence/absence matrices with 1000 observations of the variables.



We then inferred 100 simplicial complexes. The number of instances that produced specific numbers of false/true positives are shown in the following histograms for 1 and 2-simplices.



Real data

This dataset comes from the Québec Breeding Bird Atlas 2019 in which 115 species of birds have been identified on 1382 sites. From this data set, we were able to extract a simplicial complex with the following structure :

0-simplices	Totally Independent species	1-simplices	Empty triangles	2-simplices
115	4	590	917	36

We thank the Québec Breeding Bird Atlas for supplying data. We would also like to thank the following partners: Regroupement QuébecOiseaux, Environment and Climate Change Canada and Birds Canada, as well as all of the volunteer participants who gathered data for the project. We also thank the following partners for funding and resources :