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

Xavier Roy-Pomerleau, Louis J. Dubé, Patrick Desrosiers

Step 1: Fill a contingency table for each pair

<table>
<thead>
<tr>
<th>Species A</th>
<th>Species B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Step 2: Set hypotheses and corresponding log-linear models

H₀: Species i and j occur independently.

H₁: Species i and j are correlated.

Step 3: Find expected values under H₀

\[ \log(\frac{N!}{x_{ij}!x_{i}!x_{j}!}) = \sum_{i,j} x_{ij} \log(m_{ij}) - N \log(N) \]

Step 4: Test H₀ using \( \chi^2 \) statistics

\[ \chi^2 = \sum_{i,j} \frac{(x_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} \]

We reject the hypothesis with a significance level \( \alpha \) if the probability of drawing \( \chi^2 \) from a \( \chi^2 \) distribution is smaller than \( \alpha \).

Validation of the inference method with a generative model

Using a Metropolis-Hasting sampling scheme and the total distribution of the factor graph, one can generate synthetic observations.

Results on two real datasets

38 thermokarst (ponds created by the thawing of permafrost) in Northern Quebec, Canada, were sampled. The identified microorganisms were separated in 2611 taxonomic groups.

185 sites in the forests of the Côte-Nord, Quebec, Canada were sampled. 70 bird species were identified.