Joint Species Distribution Modeling: a dimension reduction approach

Duke University, *SAMSI, †North Carolina State University, ‡University of Missouri

Objective

The primary tool to jointly determine the distribution of species are Joint Species Distribution Models (JSDM). Since these account for species dependence, this is a computationally challenging problem when many species (≥100) are considered. We develop a computationally feasible approach to estimate joint species distribution models, where the number of species is of order 10^4. This approach allows:

- assessing the dependence structure across species
- identifying groups of species that share similar dependence patterns
- jointly and conditionally predict species distributions

Motivation

In the simplest possible case with continuous response (e.g., biomass), at sites \( i = 1, \ldots, n \), for \( S \) species we have

\[
y_i = B x_i + \epsilon_i \text{ with } \epsilon_i \sim N_S(0, \Sigma)
\]
where \( B_{S \times p} \) is the coefficient matrix.

- When modeling the covariance the number of parameters in these models for \( S \) species is of order \( O(S^2) \), so for large \( S \)
  \( \Rightarrow \) computationally expensive
- Many ecological applications (e.g. modeling bacterial communities) consider hundreds or thousands of species

Our strategy

Two-fold reduction:

- Approximate \( \Sigma \) with \( \hat{\Sigma} = \hat{A} \hat{\Sigma} + \sigma^2 I_S \), where \( \hat{A} \) is tall and skinny
- Identify rows of \( \hat{A} \) that can be clustered to further reduce the dimension of the problem

1. Low rank approximation of \( \Sigma \)

\[
\Sigma = AA' \text{ if } A = \Sigma^{1/2}, \text{ by introducing iid latent vectors } w_i \text{ (1) can be reformulated as}
\]

\[
y_i = B x_i + A w_i \text{ with } w_i \sim N_S(0, I_S)
\]

Now, let \( \Sigma = \hat{A} \hat{A}' + \sigma^2 I_S \approx \Sigma \), where \( \hat{A} \) is \( S \times r \) with \( r \ll S \), and the \( \sigma^2 I_S \) term yields diagonal dominance. Then, we may write

\[
y_i \approx B x_i + \hat{A} w_i + \epsilon_i \tag{3}
\]
with \( \epsilon_i \sim N_S(0, \sigma^2 I_S) \)

# parameters in covariance drops from \( S(S+1)/2 \) to \( S \times r! \)

Hierarchical formulation

\[
y_i | \hat{A}^{(k)}, B, \Sigma \sim N_S(B x_i + \hat{A}^{(k)} w_i, \sigma^2 I_S)
\]

\[
\begin{align*}
  w_i & \sim N_p(0, I_p) \\
  k_i | p & \sim \text{Dir}(p) \\
  a_i | \Omega & \sim N_p(0, \Omega) \\
  p & \sim GD(\alpha, \beta)
\end{align*}
\]

Simulations: Binary response \( S=100 \)

Data generation: \( S = 100, n = 1000, B_{r \times S} \text{ is } S \times 2, \Sigma_{true} = (\Psi \Psi')^{-1} \text{, for } \Psi = MN_{S \times S}(0, S, I_S, I_S) \)

Tuning parameter: \( r = 3 \)

Extension to Binary data

\[
R = D^{-1/2} \Sigma D^{-1/2}, \text{ with } D = \text{diag}(\Sigma)
\]

\[
y_i = (y_i^{(1)}, \ldots, y_i^{(S)}) \text{, with } y_i^{(j)} \sim \text{Bern}(\Phi(x_i^{(j)} \beta_j))
\]

\[
v_i \sim N_S(B x_i, R)
\]

Then:

\[
\Pr(y_i = q_i) = \int_{\{q_i\}} \phi_S(v_i | B x_i, R) / \nu_i \\
= \int_{\{q_i\}} \phi_S(v_i | \Sigma) / \nu_i \times \phi_S(w_i | 0, I_S) / \nu_i \times \phi_S(d w_i)
\]

with \( v_i = D^{1/2} v_i \) and \( B^* = D^{1/2} B \)

2. Clustering the rows of \( \hat{A} \)

- For row \( a_i \), introduce index \( k_i \) that denotes cluster membership, such that

\[
k_i | p \sim \text{Dir}(p) \quad \text{with } \quad p \sim GD(\alpha, \beta)
\]

where \( N(\gamma) \) is the maximum number of clusters allowed.
- Let \( S_k = \{j \in \{1, \ldots, S\} : j = k^{th}\} \)
  \( \Rightarrow \forall j \in S_k \) necessarily \( a_j = a_{k} \)
- Denote by \( \hat{A}^{(k)} \) the \( \hat{A} \) matrix conditioned on the cluster indices

Contact info & References

Email: dt108@stat.duke.edu