Concepts in model-based clustering

Francis KC Hui

Summer School on model-based multivariate analysis for ecologists

Outline

- Clustering sites and/or species
- Clustering using covariates

Questions so far?

The models so far

Throughout the course of the summer school, many of the models we have studied so far can be written in the generic form:

$$\eta_{ij} = \beta_{0j} + \alpha_i + \delta_{ij}, \tag{1}$$

where:

- β_{0j} are species-specific intercepts (column standardization);
- α_i are (optional) row effects (row standardization);
- δ_{ij} is "stuff" e.g., effects of measured covariates, latent variables, traits and phylogeny etc. . .

The models so far

Throughout the course of the summer school, many of the models we have studied so far can be written in the generic form:

$$\eta_{ij} = \beta_{0j} + \alpha_i + \delta_{ij}, \tag{2}$$

For the next little bit, we will assume α_i is always included i.e., both rows and columns are standardized.

By doing so, we can focus on the δ_{ij} part of the model i.e., what is left over after adjusting for heterogeneity in recorded species prevalence and site sampling effort.

What to do about δ_{ij}?

On Wednesday, we covered the idea of model-based ordination or some variation thereof, where $\delta_{ij} = \mathbf{u}_i^\top \gamma_j$. Provided the number of latent variables is small, then the $\mathbf{u}_i$'s and/or γ_j's can be plotted in some way to give a low-dimensional representation of patterns in species composition/indicator species etc. . .

What to do about δ_{ij}?

On Wednesday, we covered the idea of model-based ordination or some variation thereof, where $\delta_{ij} = \mathbf{u}_i^\top \gamma_j$. Provided the number of latent variables is small, then the $\mathbf{u}_i$'s and/or γ_j's can be plotted in some way to give a low-dimensional representation of patterns in species composition/indicator species etc...

In this lecture, we will talk about another way to model the δ_{ij}'s using ideas from clustering...

What to do about δ_{ij}?

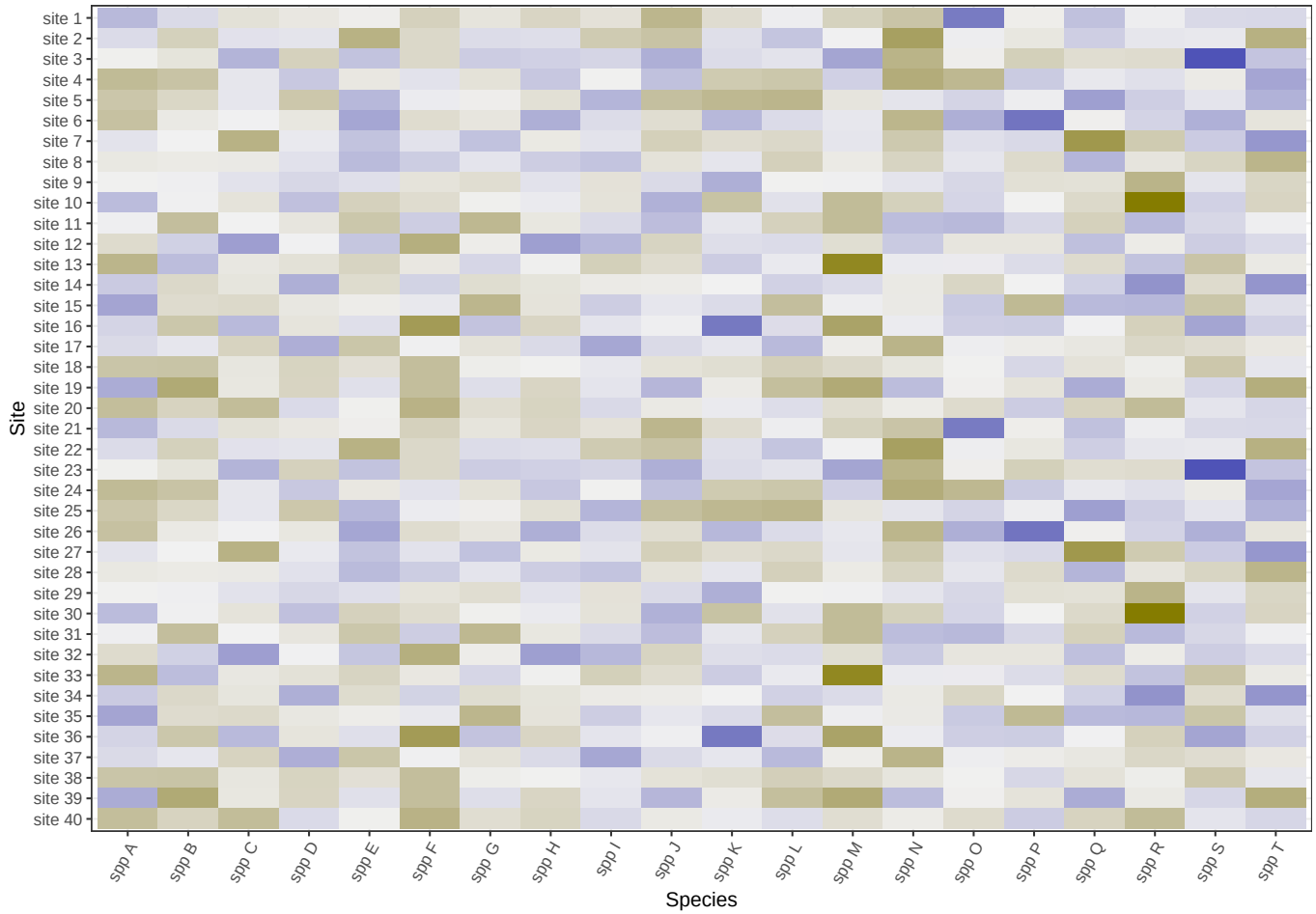
Consider again the model

$$\eta_{ij} = \beta_{0j} + \alpha_i + \delta_{ij}, \tag{3}$$

and suppose now the δ_{ij}'s are just directly estimated as fixed effects, alongside the β_{0j}'s and α_i. We will refer this as the **saturated** model, since:

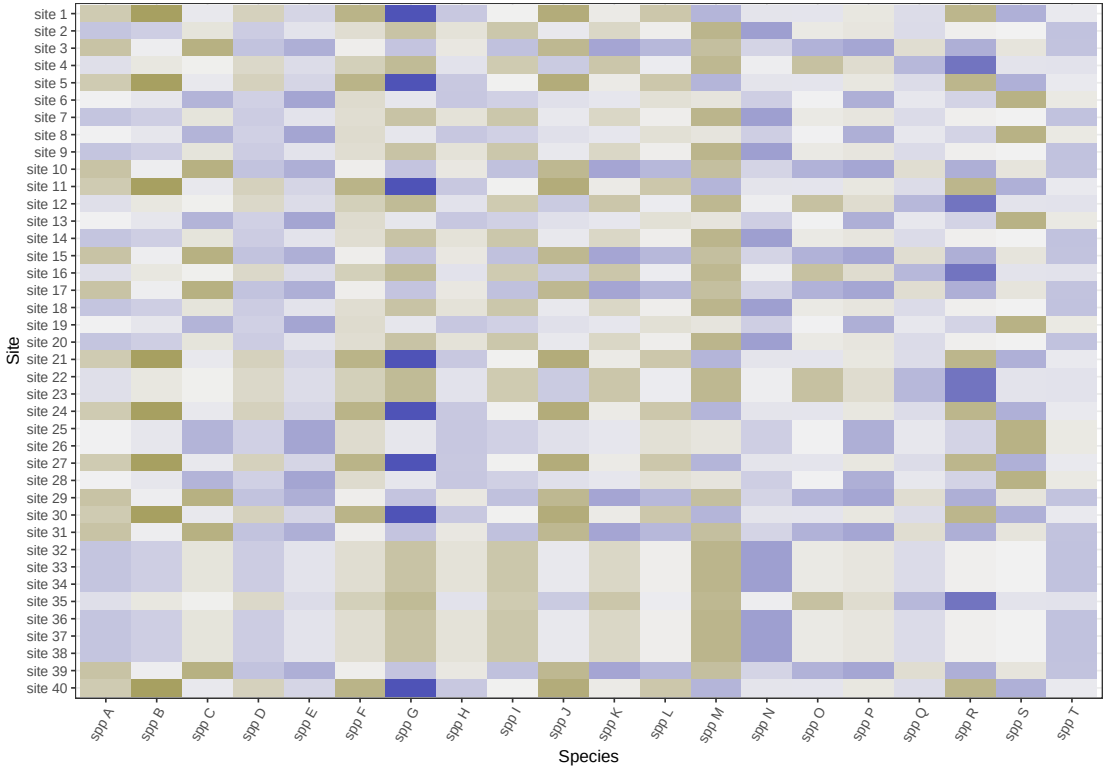
- it estimates a unique "interaction" for every combination of sites and species;
- the number of parameters is basically the same as the number of observations.

What to do about δ_{ij}?



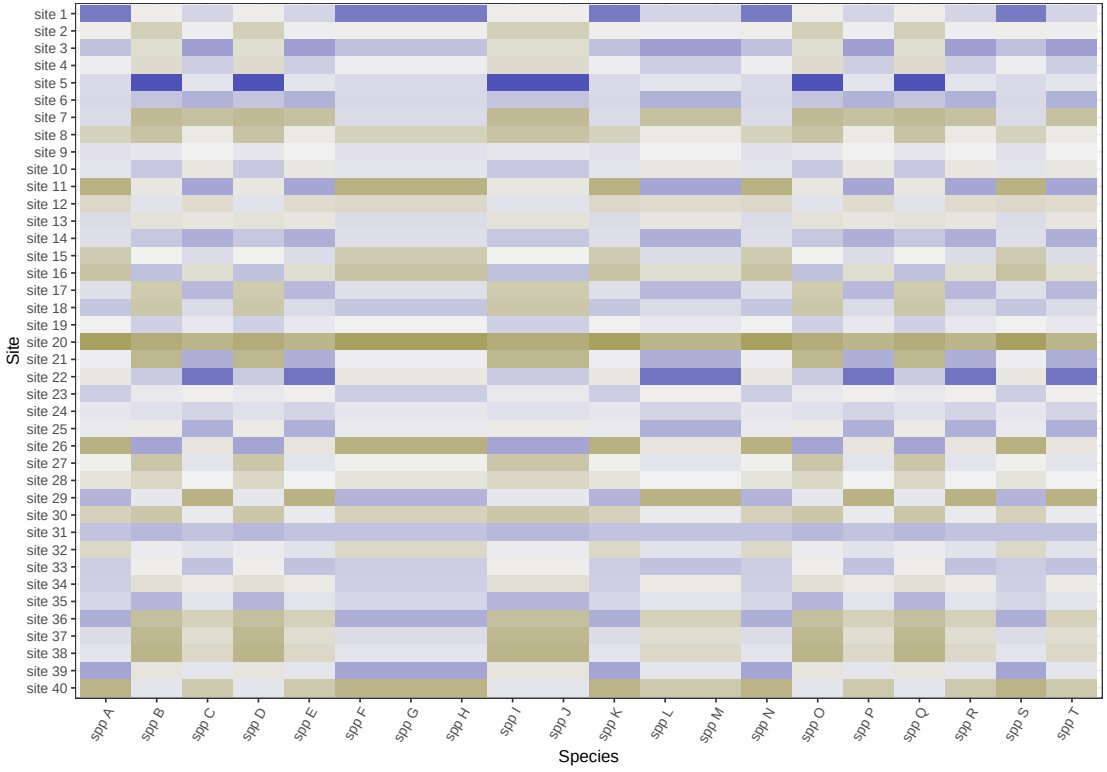
Simplifying the δ_{ij}'s

The above may be overly complex for many multivariate abundance datasets however, as in practice there may be **row patterns**



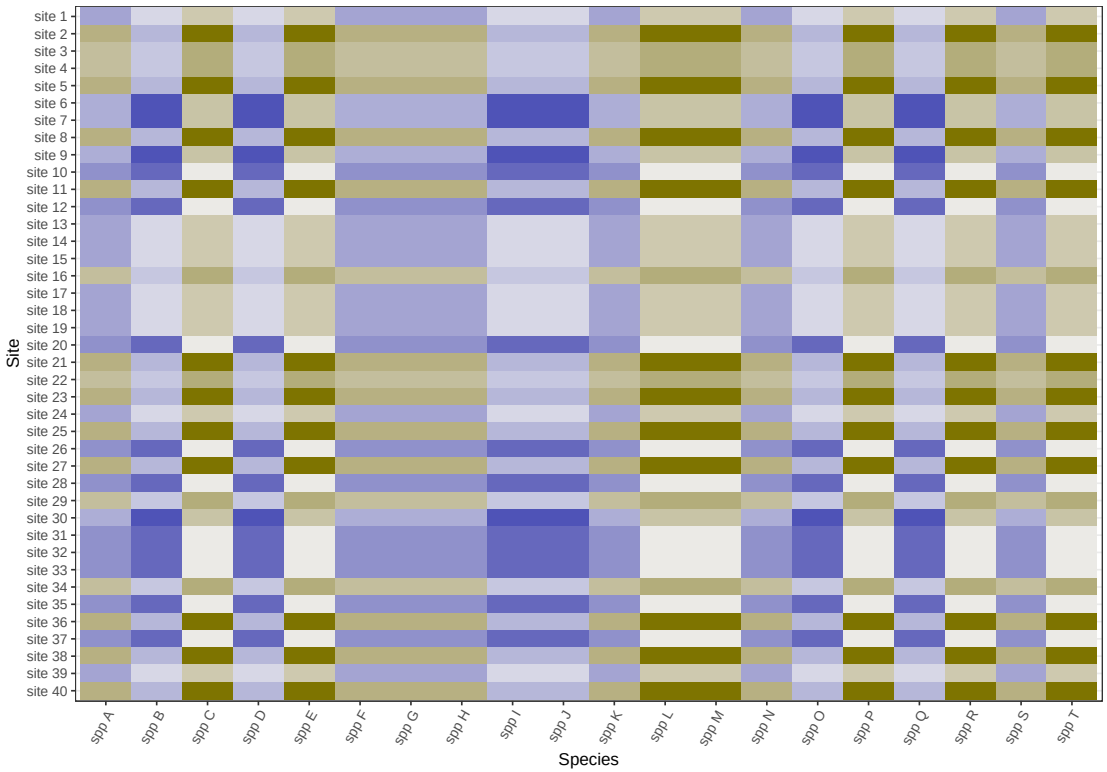
Simplifying the δ_{ij}'s

The above may be overly complex for many multivariate abundance datasets however, as in practice there may be **column patterns**



Simplifying the δ_{ij}'s

The above may be overly complex for many multivariate abundance datasets however, as in practice there may be **row & column patterns**



Clustering the δ_{ij}'s

The above motivates a way of simplifying the saturated model, namely by clustering the interaction terms δ_{ij}'s based on the row and/or column indices:

- **Row/site clustering**: δ_{rj} where $r = 1, \ldots, R < n$;
- **Column/species clustering**: δ_{ic} where $c = 1, \ldots, C < m$;
- **Biclustering**: δ_{rc}.

Row pattern detection model

Assume the patterns of site relative abundance can be in clustered into one of $R < n$ groups:

$$\eta_{ij} = \beta_{0j} + \alpha_i + \delta_{rj}. \tag{4}$$

Intuition: The assemblage is comprised of only a small number of "species profiles". Two sites i and i' in the same species profile have the same relative abundance and only differ in their α_i's e.g., site total abundance, sampling effort, and so on.

Column pattern detection model

Assume the species can be in clustered into one of $C < m$ groups:

$$\eta_{ij} = \beta_{0j} + \alpha_i + \delta_{ic}. \tag{5}$$

Intuition: The species in the assemblage can be classified into a small number of "archetypes" (or guilds?). Two species j and j' in the same guild have the same distribution across sites, and only differ in their β_{0j}'s e.g., overall prevalence.

Biclustering pattern detection model

Assume the species can be in clustered into one of $C < m$ groups:

$$\eta_{ij} = \beta_{0j} + \alpha_i + \delta_{ic}. \tag{6}$$

Intuition: Combine the ideas of species profiles and guilds