Concepts in model-based clustering

Francis KC Hui

Summer School on model-based multivariate analysis for ecologists

Outline

- Clustering sites and/or species
- Clustering incorporating 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:

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

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:

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

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}?

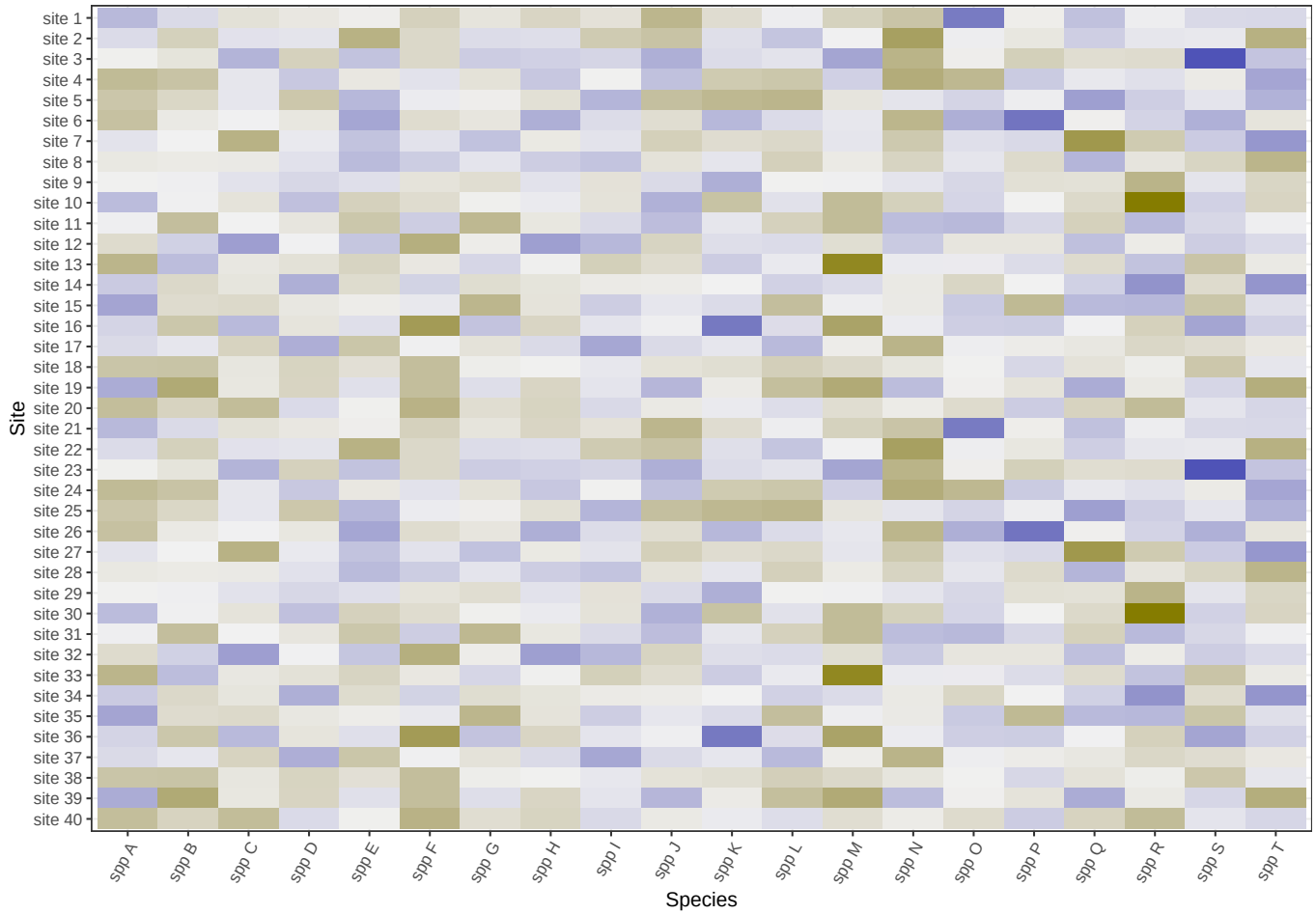
Consider again the model

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

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:

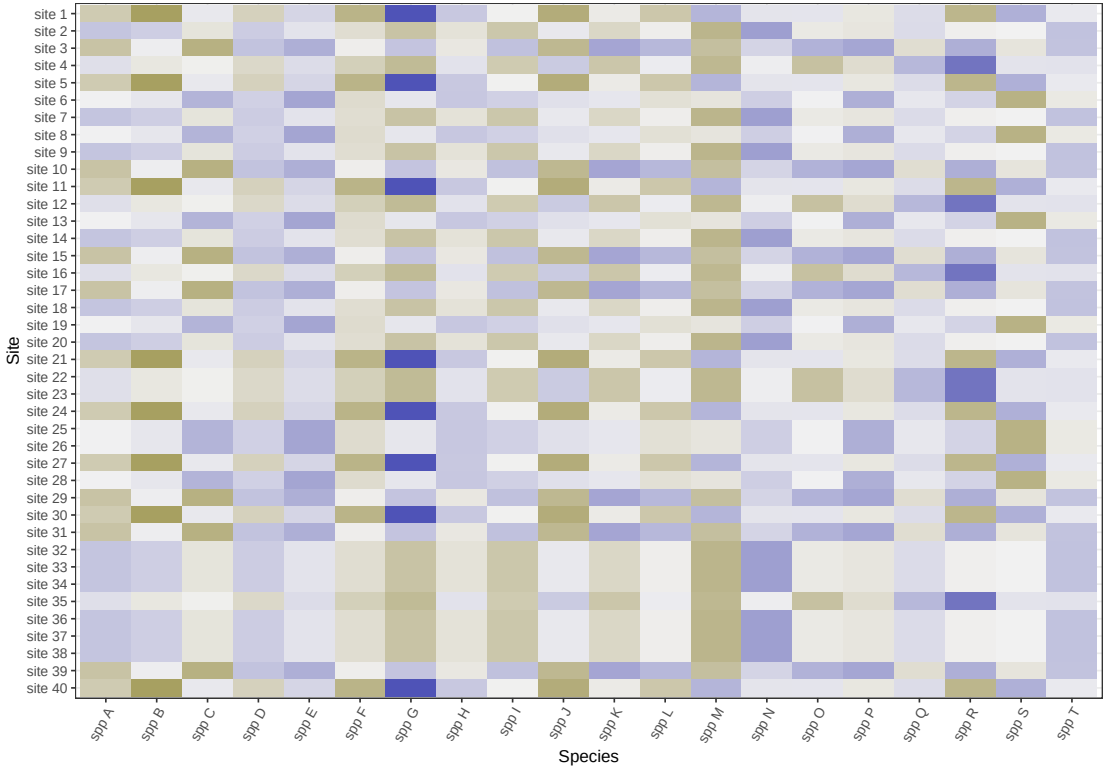
- it estimates a unique "interaction" for every combination of sites and species;
- the number of parameters is basically the same as the number of observations.

What to do about δ_{ij}?



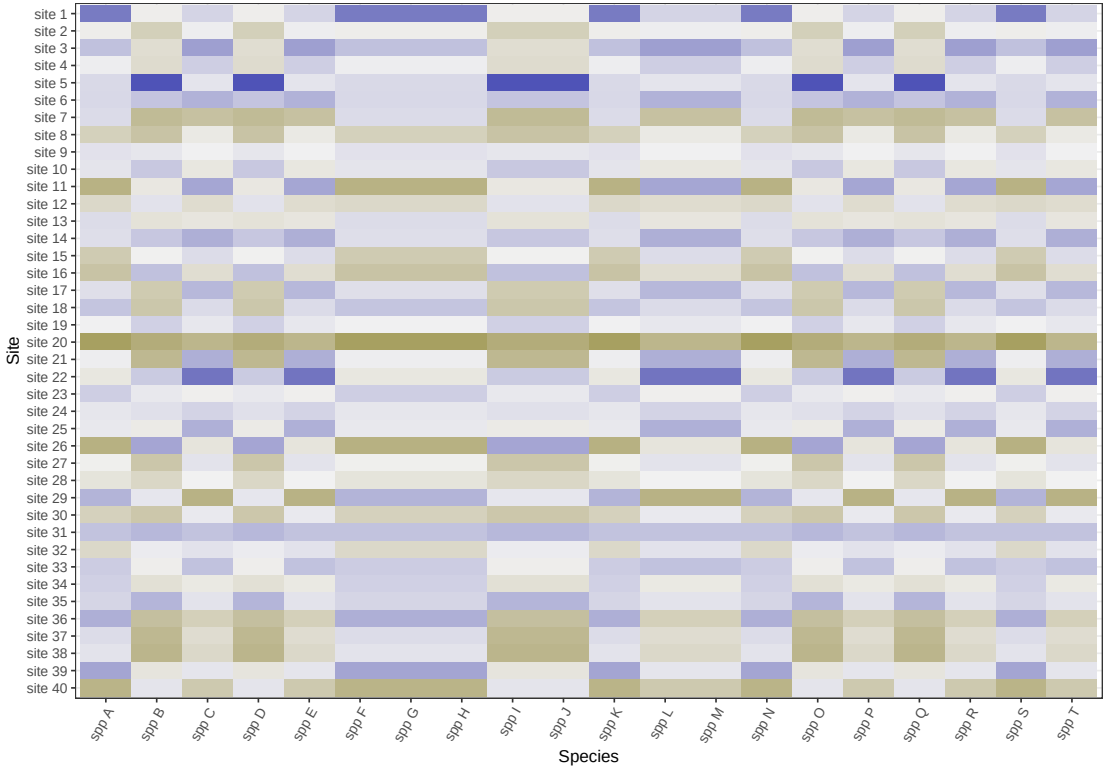
Simplifying the δ_{ij}'s

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



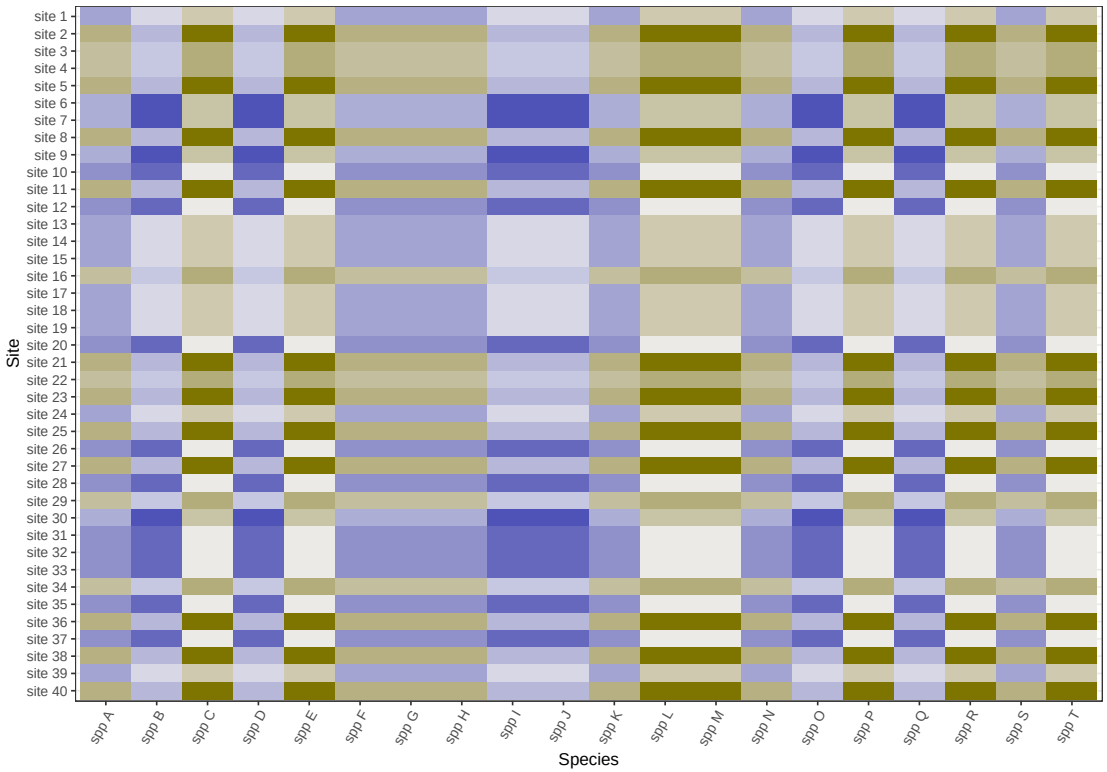
Simplifying the δ_{ij}'s

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 clustered into one of $R < n$ groups:

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

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 clustered into one of $C < m$ gr

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

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 clustered into one of $C < m$ groups:

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

Intuition: Combine the ideas of species profiles and guilds.

We will talk briefly about the stats behind these models a bit later. . .

Pattern detection models

Like model-based unconstrained ordination, clustering/pattern detection offers a parsimonious way of understanding the species community. But their goals are different:

- Is your goal a low-dimensional representation of how species composition varies over sites, or do you want to find sites with the same/similar species composition?
- Is your goal a low-dimensional representation of which species primarily drive composition across sites, or do you want to find species with the same/similar distributions?

Can read a bit more about the differing characteristics of ordination versus clustering in Section 10.1 of Legendre and Legendre, 2012, and Section 3.1 of McGarigal et al., 2000.

Pattern detection models

Note environmental covariates and species traits are not utilized in the PD models discussed so far, analogous to model-based unconstrained ordination. We will discuss how to incorporate covariates later.

Pattern detection models

Note environmental covariates and species traits are not utilized in the PD models discussed so far, analogous to model-based unconstrained ordination. We will discuss how to incorporate covariates later.

What about doing both unconstrained ordination and clustering simultaneously?

Pattern detection models

Note environmental covariates and species traits are not utilized in the PD models discussed so far, analogous to model-based unconstrained ordination. We will discuss more about how to incorporate covariates later.

What about doing both unconstrained ordination and clustering simultaneously?

You can! We won't talk about it in this lecture, but see Hui, 2017 and Stratton et al., 2024 for some ideas.

Pattern detection models – A bit of statistics

To fit something like the row pattern detection model with $R < n$,

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

we can assume:

1. The prior probability for site i to belong to cluster r is given by π_r (to be estimated);
2. The species are independent of each other, conditional on belong to the same cluster (big assumption!);
3. Species responses y_{ij} come from some distribution with mean given by $\mu_{ij} = g^{-1}(\eta_{ij})$ plus some species-specific dispersion/nuisance parameters as required

Statistically, this results in a finite mixture model.

Pattern detection models – A bit of statistics

Statistically, this results in a finite mixture model.

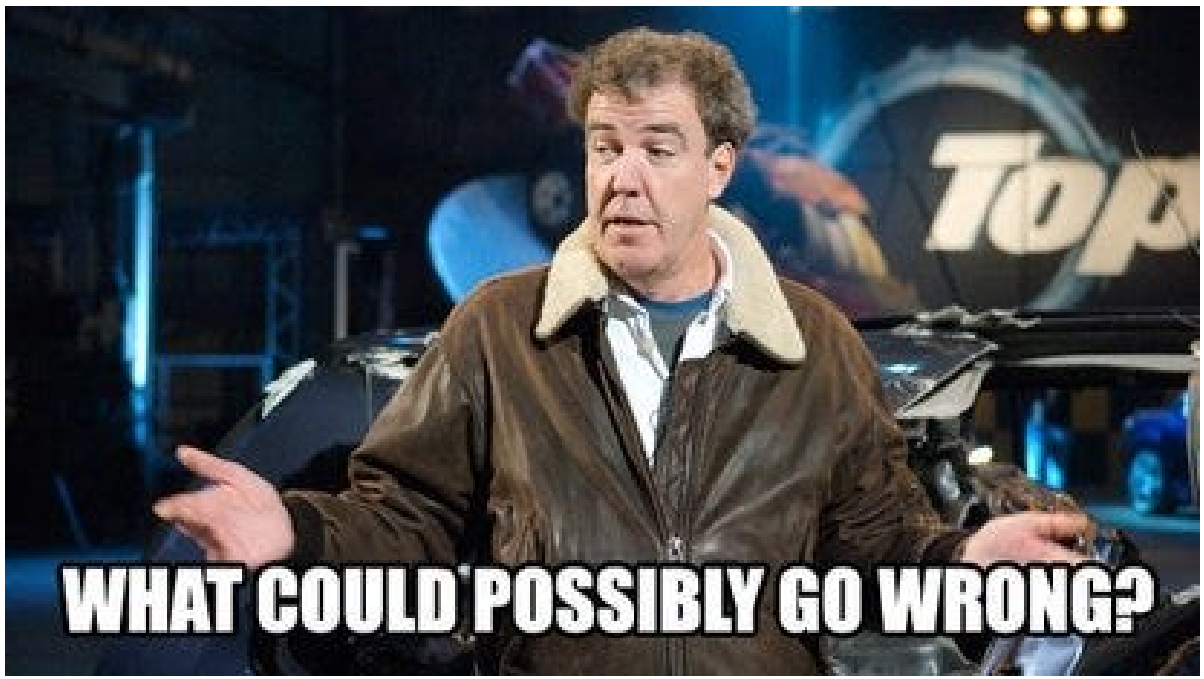
Things to think about:

- What distribution to assume for y_{ij} e.g., presence-absence, counts, biomass, percent cover etc. . .
- What to choose for R and/or C
- Do you need the row effects α_i? Depends on the interpretation. . .

Pattern detection models

Brief live demonstration using the `R` package clustglm and the `aviurba` dataset in ade4.

You can also work through the clustglmTutorial_newapplication.R script in your own time.



Pattern detection models

Brief live demonstration using the `R` package clustglm and the `aviurba` dataset in ade4.

- Can use AICs and BICs to choose the number of row/column clusters. Can also use it to decide what kind of PD model you want, but really that should be governed by the question of interest!
- Model diagnostics are possible like with other models introduced previously;
- Profile plots as a way of visualizing the uncovered cluster profiles i.e., species profiles, archetypes of guilds;
- The more mathematically curious can read Pledger and Arnold, 2014

The state of play with `clustglm`

Unfortunately, clustglm is not actively maintained, and has lots of limitations:

- Handles only a very limited number of response types;
- Standard errors are not done properly;
- Slow;
- Was an attempt, but at the moment does not practically do enough, to replace aspects of distance-based clustering

The state of play with `clustglm`

Clustering incorporating covariates

There are two (main) ways we can think of involving measured covariates $\mathbf{x}_i$ in the clustering process:

1. Cluster species that have the same/similar responses to environment/habitat factors;
2. Use environment/habitat factors to drive the process of clustering sites with similar species profiles;

Clustering incorporating covariates

There are two (main) ways we can think of involving measured covariates $\mathbf{x}_i$ in the clustering process:

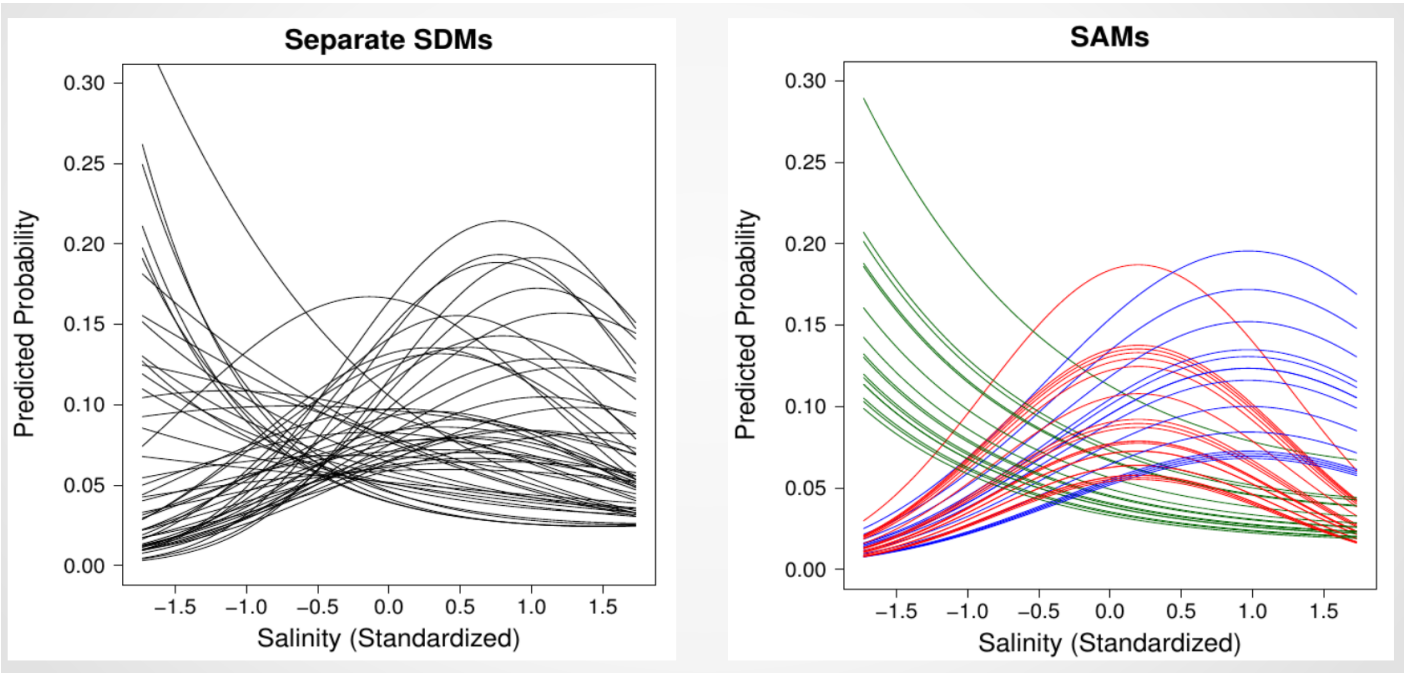
1. Cluster species that have the same/similar responses to environment/habitat factors;
2. Use environment/habitat factors to drive the process of clustering sites with similar species profiles;

The former leads to species guilds or **species archetype models**, and the latter leads to **model-based bioregionalization**

Species archetype models

Some starting papers in SAMs include Dunstan et al., 2011, Dunstan et al., 2013, and Hui et al., 2013.

But intuitively, we want to cluster on the shape of the environmental response.



Species archetype models

Statistically, and removing the optional row effects α_i as the mindset is now more about JSDMs than ordination, SAMs can be written as:

$$\eta_{ij} = \beta_{0j} + \mathbf{x}_i^\top \beta_c + \text{stuff...},$$

where species are clustered into $C < m$ archetypes, and the intercepts are specific to each species. We also assume the prior probability for species j to belong to archetype c is given by π_c (to be estimated). Note the link to column pattern detection models!

Species archetype models

Statistically, and removing the optional row effects α_i as the mindset is now more about JSDMs than ordination, SAMs can be written as:

$$\eta_{ij} = \beta_{0j} + \mathbf{x}_i^\top \beta_c + \text{stuff...},$$

where species are clustered into $C < m$ archetypes, and the intercepts are specific to each species. We also assume the prior probability for species j to belong to archetype c is given by π_c (to be estimated). Note the link to column pattern detection models!

SAMs are again a type of finite mixture model, so you can also retrieve a posterior probability of a species belonging to an archetype.

Species archetype models

Software-wise, ecomix can fit these models but is slow and has limitations. assam, by yours truly, is WIP that is more approximate but computationally more scalable and flexible.

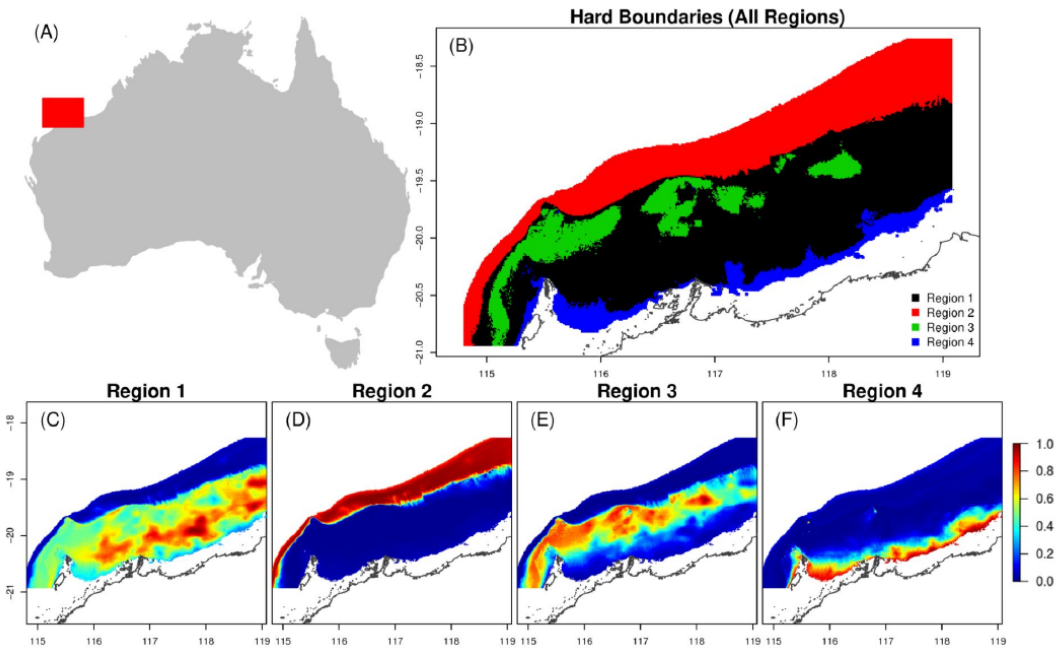
Things to think about:

- What distribution to assume for y_{ij} e.g., presence-absence, counts, biomass, percent cover etc. . .
- How many archetypes?
- What to include in stuff?

Model-based bioregionalization

Some starting papers in SAMs include Foster et al., 2013, Hill et al., 2020, and Woolley et al., 2020.

But intuitively, sites with similar environments/habitats should have similar species profiles, even if they are far apart in space.



Model-based bioregionalization

Statistically, and removing the optional row effects α_i as the mindset is now more about JSDMs than ordination, such bioregion models can be written as:

$$\eta_{ij} = \beta_{0j} + \eta_{rj} + \text{stuff}; \quad \text{P(site } i \text{ belongs to bioregion } r) = \pi_r(\mathbf{x}_i, \text{stuff2}),$$

where sites are clustered into $R < n$ bioregions. The prior probabilities $\pi_r(\mathbf{x}_i)$ can now vary with the measured covariates, and the parameters in this probability function are estimated. Note the similarity to row pattern detection models!

Model-based bioregionalization

Statistically, and removing the optional row effects α_i as the mindset is now more about JSDMs than ordination, such bioregion models can be written as:

$$\eta_{ij} = \beta_{0j} + \eta_{rj} + \text{stuff}; \quad \text{P(site } i \text{ belongs to bioregion } r) = \pi_r(\mathbf{x}_i, \text{stuff2}),$$

where sites are clustered into $R < n$ bioregions. The prior probabilities $\pi_r(\mathbf{x}_i)$ can now vary with the measured covariates, and the parameters in this probability function are estimated. Note the similarity to row pattern detection models!

Bioregion models are a type of finite mixture of experts model, so you can also retrieve a posterior probability of a site belonging to a bioregion.

Model-based bioregionalization

Software-wise, ecomix can fit these models but is slow and has limitations. Otherwise, there exists some bespoke code here and there but research is needed here!

Things to think about:

- What distribution to assume for y_{ij} e.g., presence-absence, counts, biomass, percent cover etc. . .
- How many bioregions?
- What to include in stuff and stuff2?

Outgoing remarks

- Model-based clustering ecology is much less developed than model-based ordination, both in terms of methods and software;
- Why do ecologists want to cluster?
- Model-based hierarchical clustering is a big gap!
- Conversely, models open up interesting new flavors of clustering e.g., Bystrova et al., 2021; Hui et al., 2024.