SIMULATION STUDY BACKGROUND

SARA STOUDT

1. Goals

- Is dimension reduction a realistic approach for joint species distribution modeling?
- Does the technique of dimension reduction matter?
- Can we quantify how much we sacrifice in terms of some metric if the true ecology cannot be represented well by a small number of dimensions? (TO DO: define metrics!)

2. Vocabulary

- rank of matrix - number of "linearly independent column vectors" (think of these as linear combinations of unobserved variables that affect relationships between species)
- low rank: correlation matrix can be described by a small (let's say < 5) number of "factors"
- sparse: precision matrix has many zeros in it (Note: a low rank matrix is not guaranteed to be sparse)
- dimension reduction: estimate a complicated relationship by distilling it into smaller number of variables (here species)
- Principal Component Analysis (PCA): dimension reduction technique, ordination methods often use this (I think), projection onto a lower dimensional space, focuses on the variances rather than the covariances, does not yield a sparse solution, closed form solution in normal case
- Factor Analysis: dimension reduction technique, boral (and I think to some extent hmsc) is a special case of this approach, focused on the covariances by partitioning common from unique variances, no closed form solution

3. boral model formulation

To get some insight into the latent factor approach:
site i
species j
latent factor k

Φ: CDF of standard normal

observed data: Y_{ij} true occurrence, assuming perfect detection
parameters: λ_{jk}
random effects: η_{ik}

$$
\begin{aligned}
Y_{ij} &\sim Bern(\Phi(v_{ij})) \\
v_{ij} &= \eta_{i.}\lambda'_{j.} \\
\eta_{i.} &\sim MVN(0, I_K) \\
\lambda_{jk} &\sim N(0, 10)
\end{aligned}
$$

Think of λ_{kj} as species-specific coefficients for unobserved covariates $\eta_{i.}$ at the site level. They introduce correlation between species: $v_{i.} \sim MVN(0, \Lambda\Lambda')$.

Note: ignoring fixed effects of observable covariates for now

Note: the following simulation scenarios are designed to test boral and HMSC, they may not be fair to ordination and PERMANOVA

I think this will all come down to the nuances between PCA (related to ordination and PERMANOVA) and Factor Analysis (related to boral and hmsc). Which is a better dimension reduction technique for certain cases. Which method's assumptions match the ecology better?

4. Simulation Study Regimes

4.1. **Correct Specification - Data Practicalities.**

- The model truly comes from a latent factor model. You pick the correct number of latent factors. How much data do you need to estimate the species covariance matrix well?
- Hypothesis based on relevant literature: ratio of number of species to number of sites is what matters

4.2. **Correct Form - Wrong Number of Latent Factors.**

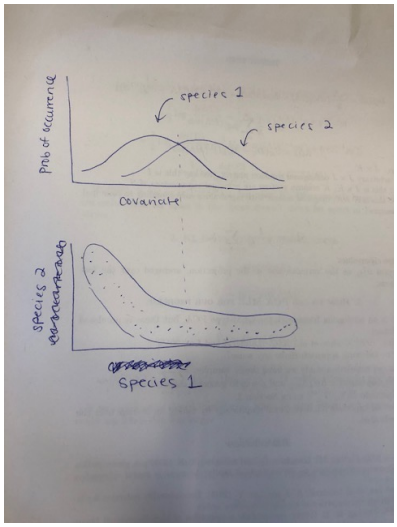
- What if nature cannot be well estimated by a small number of latent factors? How much does the act of dimension reduction hurt us?
- Hypothesis based on relevant literature: based on singular value decomposition and low rank approximations

4.3. **Latent Factors + Extra.**

- $\Lambda\Lambda'$: low rank matrix
- Σ: arbitrary precision matrix (could be sparse, medium, or dense), need to check that the rank of the inverse is large
- Perry proposed a linear combination: $a\Lambda\Lambda' + (1-a)\Sigma^{-1}$. If a is small, this is more mis-specified. A denser Σ is more mis-specified.

4.4. **Non-Linearities.**

- Perry proposed a scenario where the relationship between species is non-linear. This can be induced by having two species react to the same covariate in the same way but with a shift. If this covariate is not included in the model, then the residual covariance between them should be non-linear.



4.5. **Block Diagonal (but not too close to the identity matrix).**

- Within a group of species, there are interactions between all of them, but there are no interactions between these species and species in other groups.
- Will's hypothesis is that this type of matrix will not be well approximated by a latent factor model.

5. Instructions

- You will be given a set of datasets (rows are sites, columns are species, entries are 0 or 1 for occurrence, assuming perfect detection).
- Everyone will take ownership of one type of model.
- For each dataset, run the model and save all output (we haven't decided on metrics of evaluation yet, so who knows what we'll end up needing, better safe than sorry). These files might be big, so don't push them to GitHub. Just keep them locally until next semester.
- If your method makes you choose the dimension, fit the dataset using the true dimension d (number of latent factors, each dataset will be labeled), one fewer dimension $(d-1)$, and one larger $(d+1)$.
- If your method automatically chooses a dimension, just run every dataset once.

- Note: each dataset will represent a different simulation study regime. Eventually we'll have to run many datasets per regime, so if you want to setup your workflow to accommodate that now, that would be great.