Covariate-informed latent interaction models

Addressing geographic & taxonomic bias in predicting bird-plant interactions

Georgia Papadogeorgou

with Carolina Bello, Otso Ovaskainen & David Dunson

Motivation

- Measured networks are often
  $\rightsquigarrow$ incomplete
    we observed some interactions and recorded them
  $\rightsquigarrow$ measured with error
    incomplete + recorded interactions might not be truly present

- Incomplete measured networks might occur because researchers
  $\rightsquigarrow$ target specific individuals
  $\rightsquigarrow$ have access to only a subset of nodes

- Interest:
  $\rightsquigarrow$ Infer the true interaction network from limited measured networks
  $\rightsquigarrow$ Understand the covariates that drive node interaction

Modeling incomplete networks

- Not necessarily a problem:
  ⇝ If our inferential interest is the population we followed
- Could be a problem:
  ⇝ If the population we want to learn about is dissimilar than the one followed
  ⇝ measured interactions are not representative of interactions among target population

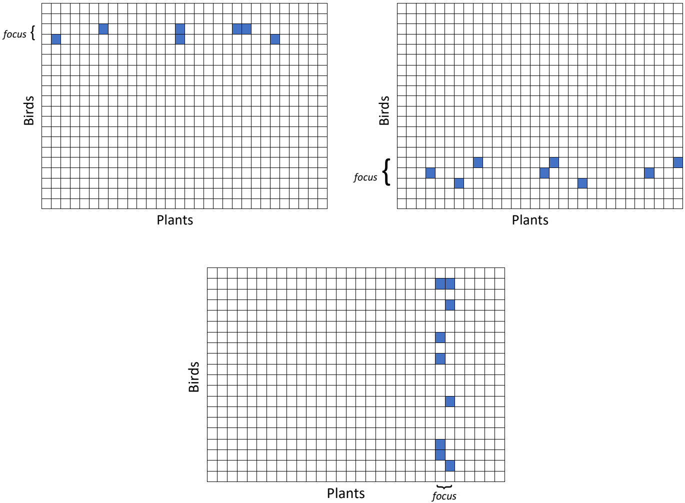
Bipartite interaction networks in ecology

- Measured networks of species interactivity are incomplete

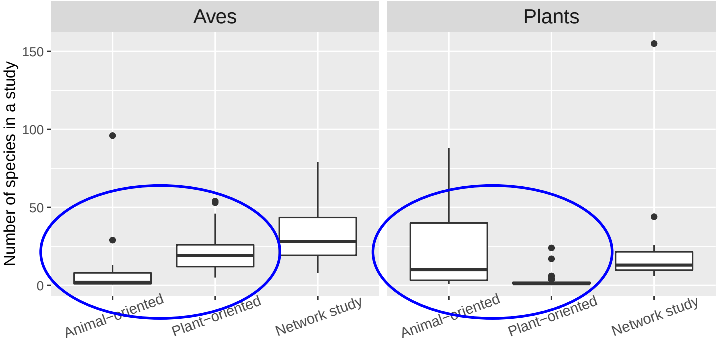
Bipartite interaction networks in ecology

- Measured networks of species interactivity are incomplete
- Individual studies on species' interactivity often focus on specific species

Bipartite interaction networks in ecology



Bipartite interaction networks in ecology

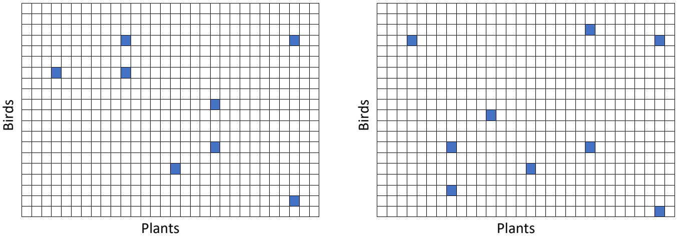


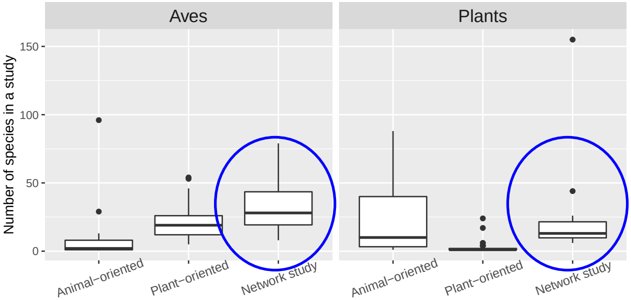
Number of unique species by study type

Bipartite interaction networks in ecology

- Measured networks of species interactivity are incomplete
- Individual studies on species' interactivity often focus on specific species
- Network studies are most useful for studying species interactions

Bipartite interaction networks in ecology

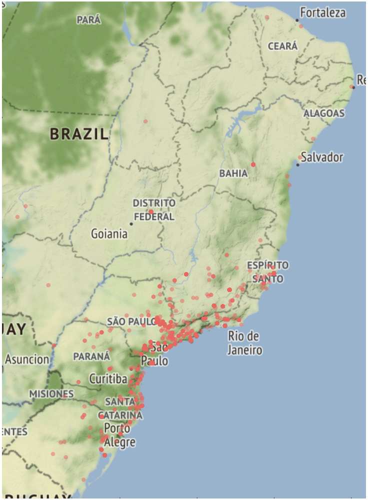




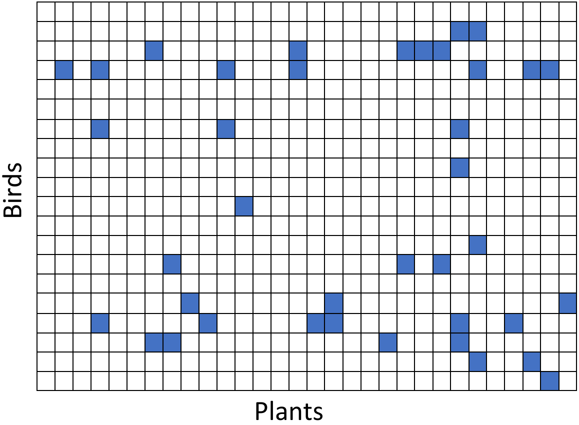
Bipartite interaction networks in ecology

- Measured networks of species interactivity are incomplete
- Individual studies on species' interactivity often focus on specific species
- Network studies are most useful for studying species interactions
- All studies focus on a specific geographic area

Bipartite interaction networks in ecology



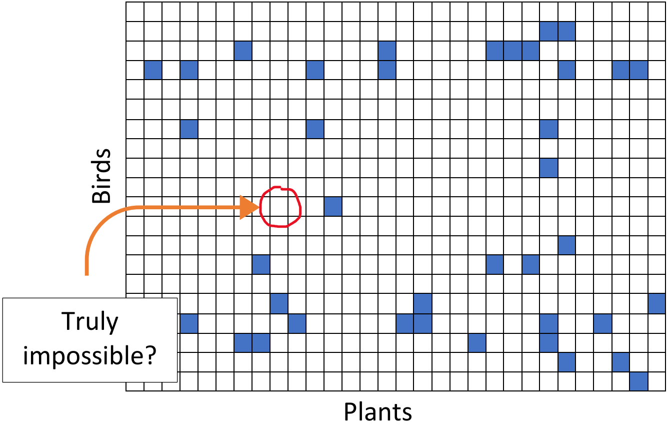
Bipartite interaction networks in ecology



Bipartite interaction networks in ecology

- Measured networks of species interactivity are incomplete
- Individual studies on species' interactivity often focus on specific species
- Network studies are most useful for studying species interactions
- All studies focus on a specific geographic area
- Under or over-representation of species
- Combined network is taxonomically and geographically biased

Bipartite interaction networks in ecology



Truly impossible?

Bipartite interaction networks in ecology

- Measured networks of species interactivity are incomplete
- Individual studies on species' interactivity often focus on specific species
- Network studies are most useful for studying species interactions
- All studies focus on a specific geographic area
- Under or over-representation of species
- Combined network is taxonomically and geographically biased

Goals:

1. Understand species interactivity while "adjusting" for these biases
2. Learn which covariates are most important in driving species interactions & detectability

Motivation

- The Atlantic Forest currently includes only 12% of its original biome
- Plants rely on frugivore populations for seed dispersal
- Reductions in frugivore populations lead to disruptions in the regeneration of ecosystems
- Climate change, reductions in natural habitats, deforestation
- How will biological communities respond?

Goals:

- Understand species interactivity
- Would a given bird consume the seed of a given plant, if given the opportunity?
- What are the drivers of species interactions?

Our setup

- $S = 85$ individual studies
  $\rightsquigarrow$ 19 animal-oriented, 45 plant-oriented, and 19 network studies
- bird $i = 1, 2, \ldots, n_B$ $\qquad (n_B = 232)$
- plant $j = 1, 2, \ldots n_P$ $\qquad (n_P = 511)$
- $A_{ijs} = 1$ or 0: recorded or unrecorded interaction in study s
- $\boldsymbol{X}_i, \boldsymbol{W}_j$: covariate information

Our setup

- $S = 85$ individual studies
  $\rightsquigarrow$ 19 animal-oriented, 45 plant-oriented, and 19 network studies
- bird $i = 1, 2, \ldots, n_B \qquad (n_B = 232)$
- plant $j = 1, 2, \ldots n_P \qquad (n_P = 511)$
- $A_{ijs} = 1$ or 0: recorded or unrecorded interaction in study s
- $\boldsymbol{X}_i, \boldsymbol{W}_j$: covariate information

Goals (in statistical terms):

- Learn bipartite network of possible interactions
  $\rightsquigarrow$ $L_{ij} = 1$ if interaction is possible, 0 otherwise
  $\rightsquigarrow$ unrecorded interactions are not necessarily impossible
  $\rightsquigarrow$ recorded networks are prone to biases
- Study covariate importance in latent network models

Our Approach

- Elucidate likelihood for $(\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{W})$

Our Approach

- Elucidate likelihood for $(\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{W})$
- The measured covariates might not include all relevant information

  Introduce latent factors:

  $\boldsymbol{U}_i = (U_{i1}, \ldots, U_{iH})^T$ for bird species
  $\boldsymbol{V}_j = (V_{j1}, \ldots, V_{jH})^T$ for plant species

  $\rightsquigarrow$ Representation of species covariate information
  $\rightsquigarrow$ Arbitrarily close to species' measured covariates

Our Approach

- Elucidate likelihood for $(\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{W})$
- For a measured network to have recorded a given interaction, all of the following need to happen:
    - species co-occur
    - researchers are interested in the pair of species
    - species truly interact
    - an interaction was detected

Our Approach

- Elucidate likelihood for $(\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{W})$
- Species occurrence: $O_{ijs} = 1$ if i, j both occur at the study site

  $\rightsquigarrow$ important for addressing geographical bias

  $\rightsquigarrow$ fixed here

Our Approach

- Elucidate likelihood for $(\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{W})$
- Species occurrence: $O_{ijs} = 1$ if i, j both occur at the study site

  $\rightsquigarrow$ important for addressing geographical bias
  $\rightsquigarrow$ fixed here
- Study focus: $F_{ijs} = 1$ if ij-pair includes focal species

  $\rightsquigarrow$ important for addressing taxonomical bias

Our Approach

- Elucidate likelihood for $(\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{W})$
- Species occurrence: $O_{ijs} = 1$ if i, j both occur at the study site

  $\rightsquigarrow$ important for addressing geographical bias
  $\rightsquigarrow$ fixed here
- Study focus: $F_{ijs} = 1$ if ij-pair includes focal species

  $\rightsquigarrow$ important for addressing taxonomical bias
- True interactions: $L_{ijs} = 1$ if ij-pair is possible to interact

Our Approach

- Elucidate likelihood for $(\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{W})$
- Species occurrence: $O_{ijs} = 1$ if i, j both occur at the study site

  $\rightsquigarrow$ important for addressing geographical bias
  $\rightsquigarrow$ fixed here
- Study focus: $F_{ijs} = 1$ if ij-pair includes focal species

  $\rightsquigarrow$ important for addressing taxonomical bias
- True interactions: $L_{ijs} = 1$ if ij-pair is possible to interact
- Species' detectability: p_i, q_j for bird i and plant j

Our Approach – Dependencies

- Focus on

$$P\big(\boldsymbol{A} = \boldsymbol{a} \mid \boldsymbol{L}, \boldsymbol{F}, \boldsymbol{O}, \{p\}, \{q\}, \{\boldsymbol{U}\}, \{\boldsymbol{V}\}, \{\boldsymbol{X}\}, \{\boldsymbol{W}\}\big)$$

Our Approach – Dependencies

- Focus on

$$P\big(\boldsymbol{A} = \boldsymbol{a} \mid \boldsymbol{L}, \boldsymbol{F}, \boldsymbol{O}, \{p\}, \{q\}, \{\boldsymbol{U}\}, \{\boldsymbol{V}\}, \{\boldsymbol{X}\}, \{\boldsymbol{W}\}\big)$$

- Dependencies across measured networks

  $\rightsquigarrow$ geographic proximity (O_{ijs})
  $\rightsquigarrow$ study focus (F_{ijs})
  $\rightsquigarrow$ truly impossible interactions (L_{ij})
  $\rightsquigarrow$ species detectability, $p_i, q_j \in (0, 1)$

Our Approach – Dependencies

- Focus on

$$P\Big(\boldsymbol{A} = \boldsymbol{a} \mid \boldsymbol{L}, \boldsymbol{F}, \boldsymbol{O}, \{p\}, \{q\}, \{\boldsymbol{U}\}, \{\boldsymbol{V}\}, \{\boldsymbol{X}\}, \{\boldsymbol{W}\}\Big)$$

- Dependencies across measured networks

  $\leadsto$ geographic proximity (O_{ijs})
  $\leadsto$ study focus (F_{ijs})
  $\leadsto$ truly impossible interactions (L_{ij})
  $\leadsto$ species detectability, $p_i, q_j \in (0, 1)$

- We can write it as

$$\prod_{i,j,s} P\left(A_{ijs} = a_{ijs} \mid L_{ij}, F_{ijs}, O_{ijs}, p_i, q_j\right)$$

Our Approach – Dependencies

- Focus on

$$P\big(\boldsymbol{A} = \boldsymbol{a} \mid \boldsymbol{L}, \boldsymbol{F}, \boldsymbol{O}, \{p\}, \{q\}, \{\boldsymbol{U}\}, \{\boldsymbol{V}\}, \{\boldsymbol{X}\}, \{\boldsymbol{W}\}\big)$$

- Dependencies across measured networks

  $\rightsquigarrow$ geographic proximity (O_{ijs})
  $\rightsquigarrow$ study focus (F_{ijs})
  $\rightsquigarrow$ truly impossible interactions (L_{ij})
  $\rightsquigarrow$ species detectability, $p_i, q_j \in (0,1)$

- We can write it as

$$\prod_{i,j,s} P\left(A_{ijs} = a_{ijs} \mid L_{ij}, F_{ijs}, O_{ijs}, p_i, q_j\right)$$

- and specify:

$$P(A_{ijs} = 1 \mid L_{ij} = l, F_{ijs} = f, O_{ijs} = o, p_i, q_j) = \begin{cases} 0, & \text{if} \quad lfo = 0 \\ p_i q_j, & \text{if} \quad lfo = 1 \end{cases}$$

Our Approach – Dependencies

- Focus on

$$P\Big(\boldsymbol{A} = \boldsymbol{a} \mid \boldsymbol{L}, \boldsymbol{F}, \boldsymbol{O}, \{p\}, \{q\}, \{\boldsymbol{U}\}, \{\boldsymbol{V}\}, \{\boldsymbol{X}\}, \{\boldsymbol{W}\}\Big)$$

- Dependencies across measured networks

  $\rightsquigarrow$ geographic proximity (O_{ijs})
  $\rightsquigarrow$ study focus (F_{ijs})
  $\rightsquigarrow$ truly impossible interactions (L_{ij})
  $\rightsquigarrow$ species detectability, $p_i, q_j \in (0,1)$

- We can write it as

$$\prod_{i,j,s} P\left(A_{ijs} = a_{ijs} \mid L_{ij}, F_{ijs}, O_{ijs}, p_i, q_j\right)$$

$$= \prod_{\substack{i,j,s \\ F_{ijs}O_{ijs}L_{ij}=1}} (p_i q_j)^{a_{ijs}} (1 - p_i q_j)^{1-a_{ijs}} \prod_{\substack{i,j,s \\ F_{ijs}O_{ijs}L_{ij}=0}} I\left(a_{ijs} = 0\right)$$

Our Approach

- Need to specify joint distribution on unobserved quantities:

$$\boldsymbol{L}, \{\boldsymbol{U}\}, \{\boldsymbol{V}\}, \{p\}, \{q\}$$

- Still need distribution on measured covariates $\{\boldsymbol{X}\}, \{\boldsymbol{W}\}$!

Our Approach

- Need to specify joint distribution on unobserved quantities:

$$\boldsymbol{L}, \{\boldsymbol{U}\}, \{\boldsymbol{V}\}, \{p\}, \{q\}$$

- Still need distribution on measured covariates $\{\boldsymbol{X}\}, \{\boldsymbol{W}\}$!

$\rightsquigarrow$ Recorded interactions depend on species' characteristics:

$$\mathrm{logit} P(L_{ij} = 1 \mid \boldsymbol{X}_i, \boldsymbol{U}_i, \boldsymbol{W}_j, \boldsymbol{V}_j) = \lambda_0 + \sum_{h=1}^{H} \lambda_h U_{ih} V_{jh}$$

Our Approach

- Need to specify joint distribution on unobserved quantities:

$$\boldsymbol{L}, \{\boldsymbol{U}\}, \{\boldsymbol{V}\}, \{p\}, \{q\}$$

- Still need distribution on measured covariates $\{\boldsymbol{X}\}, \{\boldsymbol{W}\}$!

⇝ Recorded interactions depend on species' characteristics:

$$\mathrm{logit} P(L_{ij} = 1 \mid \boldsymbol{X}_i, \boldsymbol{U}_i, \boldsymbol{W}_j, \boldsymbol{V}_j) = \lambda_0 + \sum_{h=1}^{H} \lambda_h U_{ih} V_{jh}$$

⇝ Species' detectability depends on species' characteristics:

$$E[\mathsf{logit}(p_i) \mid \boldsymbol{U}_i, \boldsymbol{X}_i] = \delta_0 + \boldsymbol{U}_i^T \boldsymbol{\delta}$$

$$E[\mathsf{logit}(q_j) \mid \boldsymbol{V}_j, \boldsymbol{W}_j] = \zeta_0 + \boldsymbol{V}_j^T \boldsymbol{\zeta}$$

Our Approach

- Need to specify joint distribution on unobserved quantities:

$$\boldsymbol{L}, \{\boldsymbol{U}\}, \{\boldsymbol{V}\}, \{p\}, \{q\}$$

- Still need distribution on measured covariates $\{\boldsymbol{X}\}, \{\boldsymbol{W}\}$!

⤳ Recorded interactions depend on species' characteristics:

$$\operatorname{logit} P(L_{ij} = 1 \mid \boldsymbol{X}_i, \boldsymbol{U}_i, \boldsymbol{W}_j, \boldsymbol{V}_j) = \lambda_0 + \sum_{h=1}^{H} \lambda_h U_{ih} V_{jh}$$

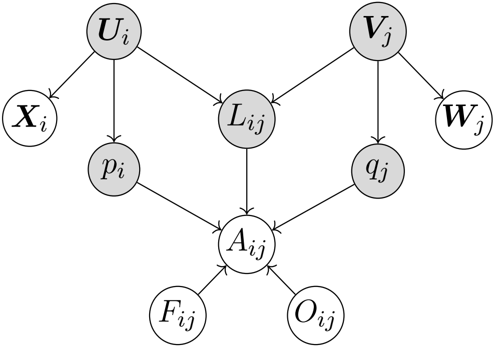
⤳ Species' detectability depends on species' characteristics:

$$E[\mathsf{logit}(p_i) \mid \boldsymbol{U}_i, \boldsymbol{X}_i] = \delta_0 + \boldsymbol{U}_i^T \boldsymbol{\delta}$$
$$E[\mathsf{logit}(q_j) \mid \boldsymbol{V}_j, \boldsymbol{W}_j] = \zeta_0 + \boldsymbol{V}_j^T \boldsymbol{\zeta}$$

⤳ Latent factors are "close" enough to measured covariates

$$f_m^{-1}(E(X_{im} \mid \boldsymbol{U}_i)) = \beta_{m0} + \boldsymbol{U}_i' \boldsymbol{\beta}_m, \ m = 1, 2, \ldots, p_B, \text{ and}$$
$$g_l^{-1}(E(W_{jl} \mid \boldsymbol{V}_j)) = \gamma_{l0} + \boldsymbol{V}_j' \boldsymbol{\gamma}_l, \ l = 1, 2, \ldots, p_P$$

Our approach (schematically)



Our Approach (prior distributions)

- $\boldsymbol{U}_{.h} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_U)$, and $\boldsymbol{V}_{.h} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_V)$
  - $\rightsquigarrow$ independently across h
  - $\rightsquigarrow$ $\boldsymbol{\Sigma}_U, \boldsymbol{\Sigma}_V$ phylogenetically structured across species

Sirio Legramanti, Daniele Durante, and David B. Dunson. Bayesian cumulative shrinkage for infinite factorizations. Biometrika, 107(3): 745â752, 2020

Our Approach (prior distributions)

- $\boldsymbol{U}_{.h} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_U)$, and $\boldsymbol{V}_{.h} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_V)$

  $\rightsquigarrow$ independently across h

  $\rightsquigarrow$ $\boldsymbol{\Sigma}_U, \boldsymbol{\Sigma}_V$ phylogenetically structured across species

---

- Model coefficients: $\beta_{mh} | \tau^{\beta}_{mh}, \theta_h \sim N(0, \tau^{\beta}_{mh} \theta_h)$ *similarly for others*

  θ_h: The increasing shrinkage prior of Legramanti et al. (2020)

  $\rightsquigarrow$ Increasingly penalizes coefficients with larger h

  τ^{β}_{mh}: Coefficient-specific additional variation

Sirio Legramanti, Daniele Durante, and David B. Dunson. Bayesian cumulative shrinkage for infinite factorizations. Biometrika, 107(3): 745â752, 2020

Computation

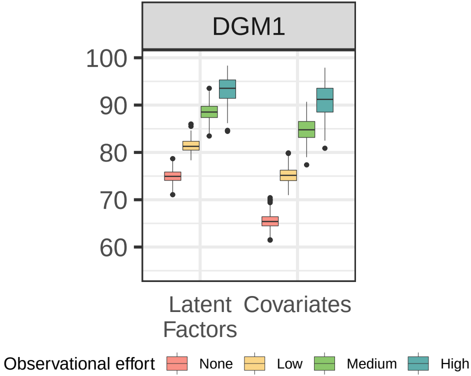
- We approximated the posterior distribution using MCMC
- Most updates were performed using Gibbs / MH
- Pólya-Gamma data augmentation scheme for parameters of logistic models

Simulations

- We considered 24 scenarios:
  - Same or different covariates important for interactions and detectability
  - Important covariates are measured, mixed or unmeasured
  - Covariates are correlated or not
  - High and low information scenarios
- Alternative approaches using covariates, ignoring biases, fixed latent factor dimension ...

Simulations

AUROC – predicting missing interactions



Variable importance in latent network models

- Interaction model does not include covariates
- We cannot interpret coefficients

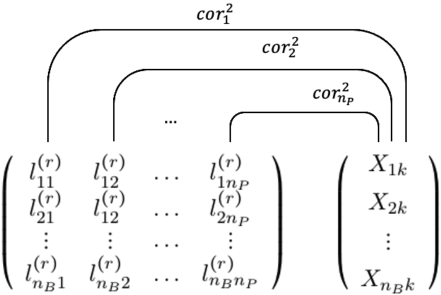
  $\rightsquigarrow$ lack of identifiability of latent factors
- Covariates are not included in the interaction model

  $\rightsquigarrow$ by design

  $\rightsquigarrow$ interpreting coefficients in models with structured latent factors has challenges (Van Ee et al., 2021)

Justin J Van Ee, Jacob S Ivan, Colorado Parks, Wildlife Mevin, B Hooten, and Mevin B Hooten. Community Confounding In Joint Species Distribution Models. 2021.

Variable importance in latent network models

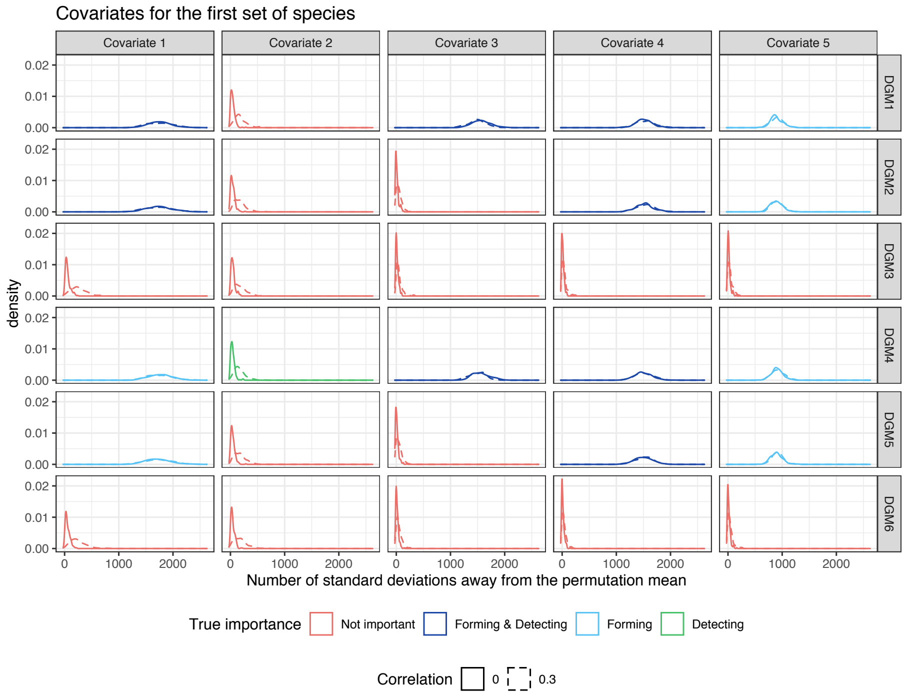

cor_1^2
cor_2^2
$cor_{n_P}^2$
...
$\begin{pmatrix} l_{11}^{(r)} & l_{12}^{(r)} & \dots & l_{1n_P}^{(r)} \\ l_{21}^{(r)} & l_{12}^{(r)} & \dots & l_{2n_P}^{(r)} \\ \vdots & \vdots & \dots & \vdots \\ l_{n_B1}^{(r)} & l_{n_B2}^{(r)} & \dots & l_{n_Bn_P}^{(r)} \end{pmatrix}$
$\begin{pmatrix} X_{1k} \\ X_{2k} \\ \vdots \\ X_{n_Bk} \end{pmatrix}$

$l_{ij}^{(r)}$: r^{th} posterior sample of (logit) probability of interaction between bird i and plant j

1. Calculate T^{obs} by averaging across species and posterior samples
2. Permute the covariate vector B times $\rightsquigarrow T^{(b)}$, $b = 1, 2, \dots, B$
3. Use

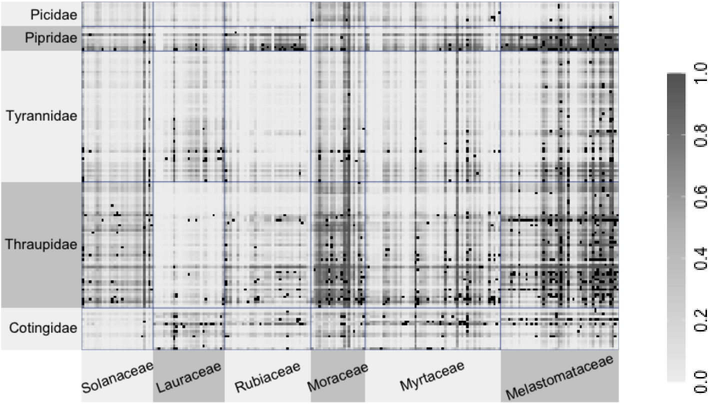
$$T^{obs} - \text{avg}\left(T^{(b)}\right) \quad \Big/ \quad \text{sd}\left(T^{(b)}\right)$$

as the variable importance metric



Covariates for the first set of species

In our study...

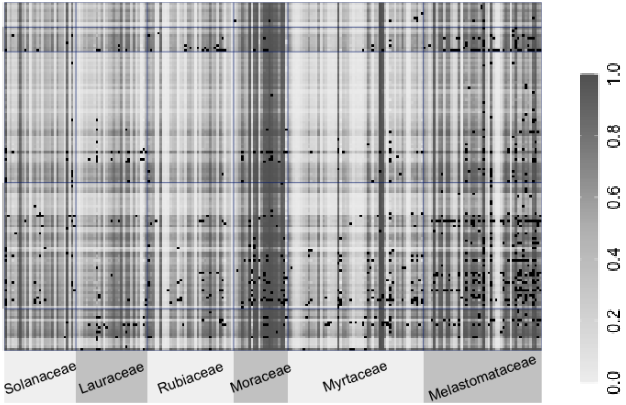


Our approach

In our study...

- Species' interaction profiles appear to be taxonomically structured
- 5% of pairs are predicted to interact (post. prob. $> 80\%$)
- 41% of pairs are predicted to *not* interact (post. prob. $< 10\%$)

In our study...



Alternative approach using covariates directly

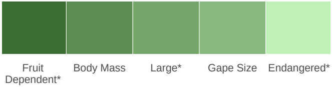
In our study...

- Species' interaction profiles appear to be taxonomically structured
- 5% of pairs are predicted to interact (post. prob. $> 80\%$)
- 41% of pairs are predicted to *not* interact (post. prob. $< 10\%$)
- Cross validation:

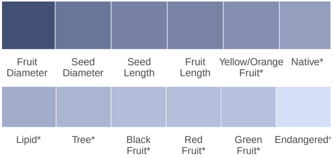
  $\rightsquigarrow$ compare post. prob. of interaction in held-out pairs compared to all pairs

  $\rightsquigarrow$ covariates: (post.prob. interacting) 1.4 times higher (all)
  $\rightsquigarrow$ latent factors: (post.prob. interacting) 3.2 times higher (all)
- Latent factor model differentiates truly possible interactions better

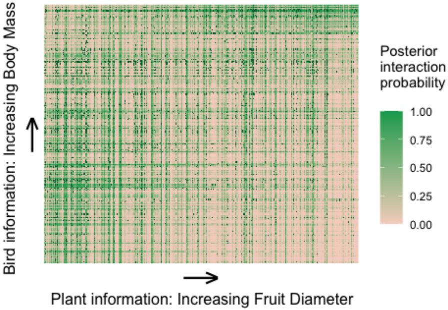
Variable importance on species interactions



(a) Bird Traits Importance



(b) Plant Traits Importance



(c) Interaction matrix ordered by traits

Final thoughts

- Latent network models for noisy bipartite networks

  ⇝ covariates inform the latent factors via separate models
  ⇝ quantifies our uncertainty around the estimated graph
  ⇝ posterior samples + permutation for variable importance

- Study species interactions based on meta-analysis data set

  ⇝ complete the bipartite graph of species interdependence
  ⇝ incorporates the missingness mechanism caused by the taxonomic and geographic bias of individual studies

- EXTENSION: simultaneous modeling of co-occurrence and interactions

  ⇝ incorporate geographic information and environmental variables
  ⇝ investigate the importance of species abundance and competition

---

Preprint:	arXiv:2103.05557
Rpackage:	`https://github.com/gpapadog/BiasedNetwork`
Analysis:	`https://github.com/gpapadog/Bird_Plant_Interactions`

Thank you!