Homogeneity pursuit and variable selection for multivariate abundance data

Francis K.C. Hui ANU
Luca Maestrini ANU
A.H. Welsh ANU

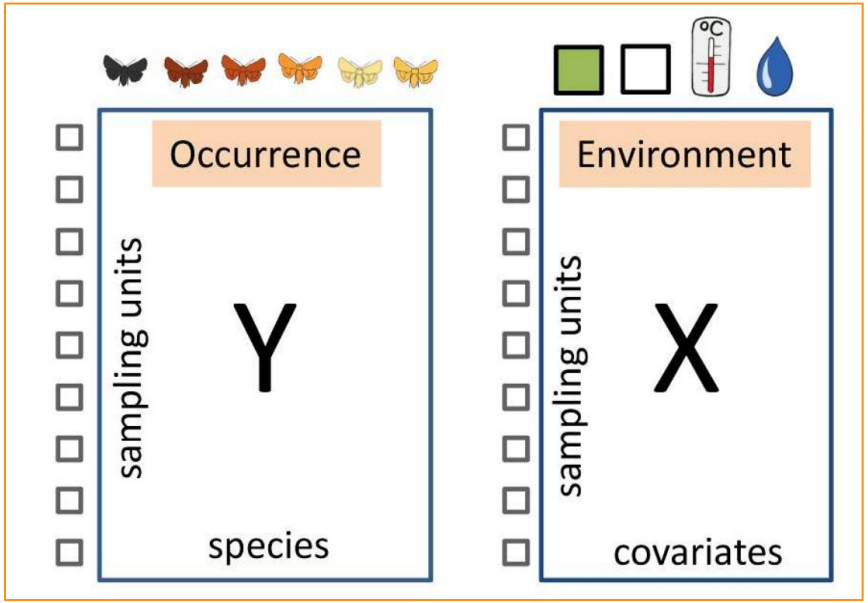
Multivariate abundance data
GEEs with some rank-reduction
HPGEE
Application to Great Barrier Reef biodiversity

Some images courtesy of SlidesCarnival and Google images
IBSAR & SEEM 2023

Multivariate abundance data

- Data characterized by:
  - Multiple, correlated species
  - Sparse, non-continuous responses

- Goal is to uncover species-environment relationships while accounting for between-species covariation

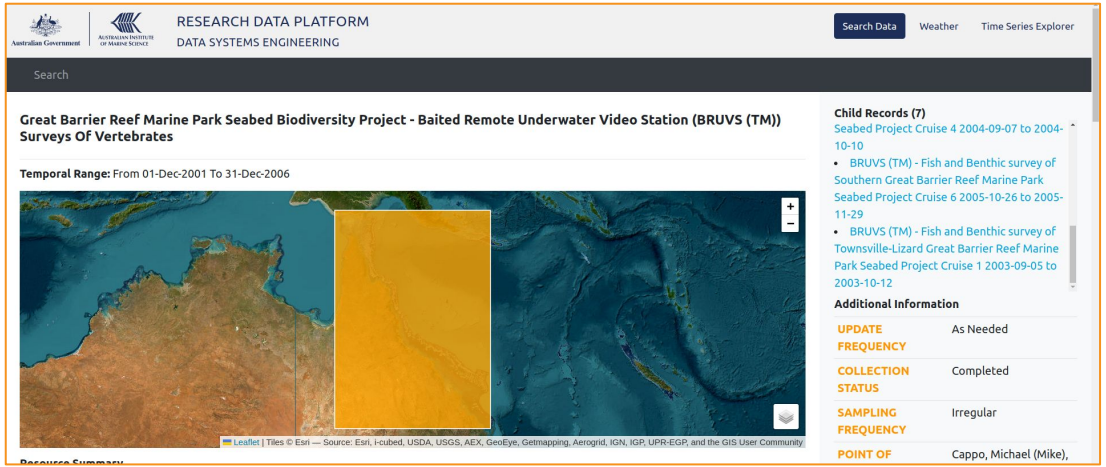


Multivariate abundance data

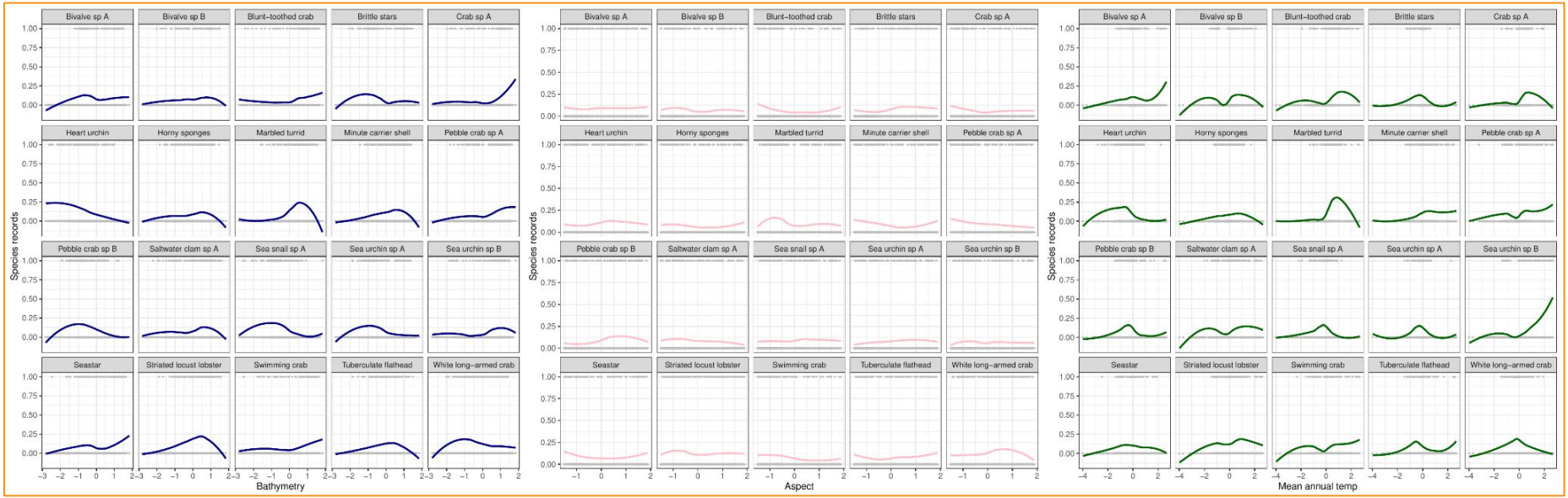
- Data characterized by:
  - Multiple, correlated species
  - Sparse, non-continuous responses

- Goal is to uncover species-environment relationships while accounting for between-species covariation

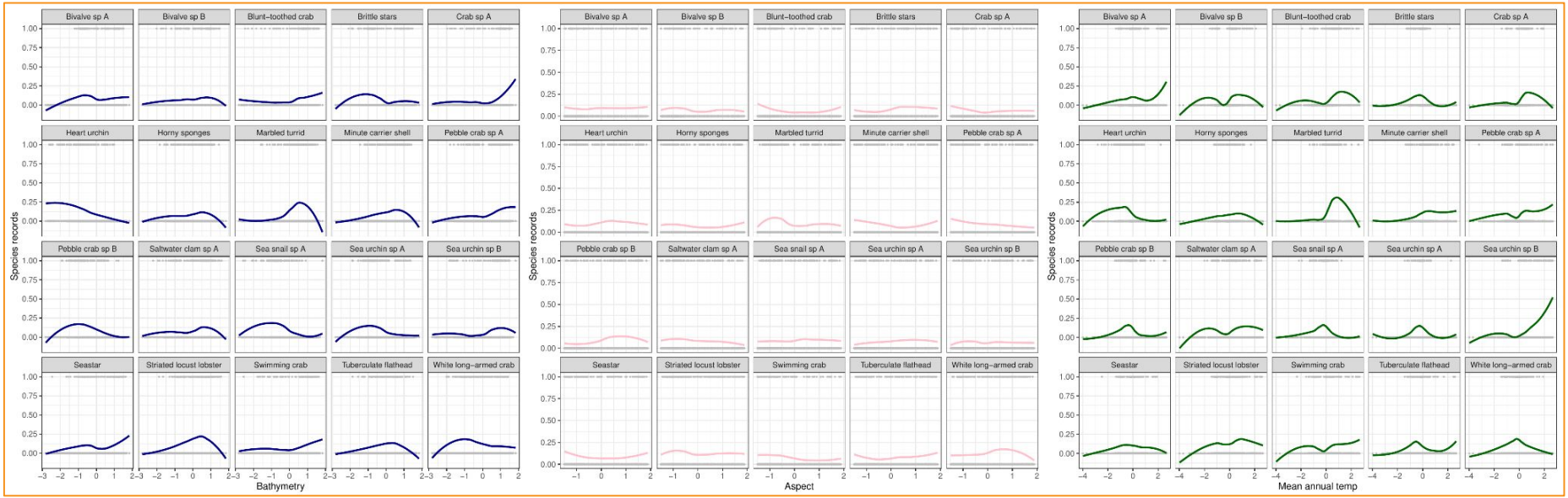
- Example from [Great Barrier Reef Seabed Biodiversity project]



Multivariate abundance data

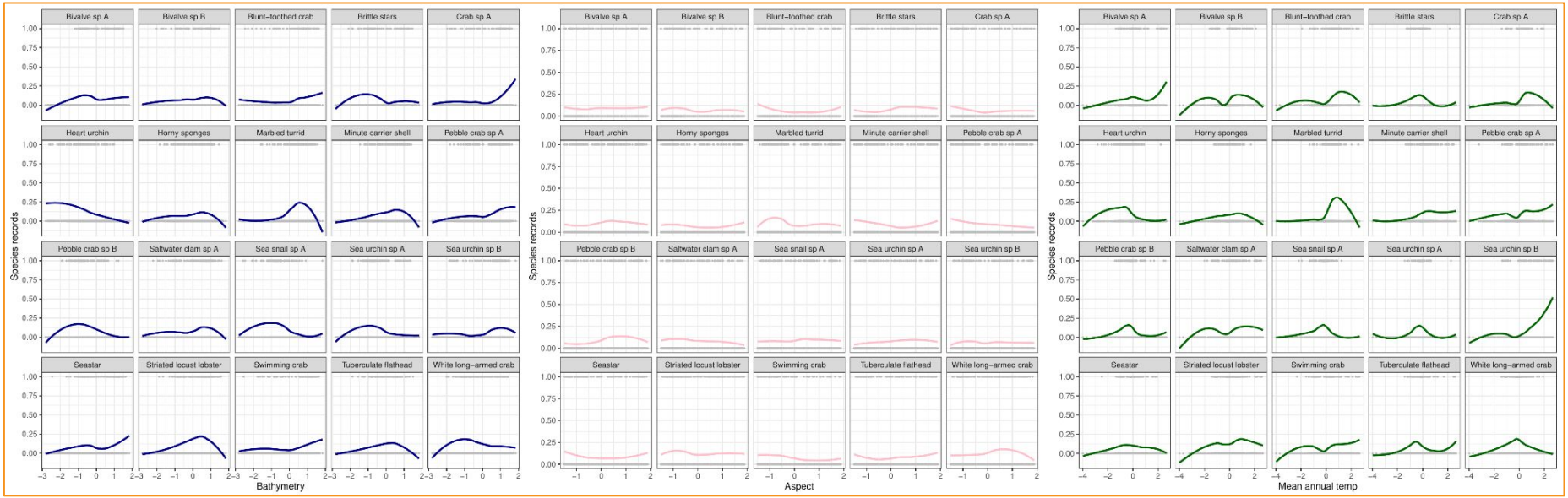


Multivariate abundance data



Species respond to the environment in different ways, informed by different subsets of covariates

Multivariate abundance data



Species respond to the environment in different ways, informed by different subsets of covariates

Within a covariate, there is homogeneity in the way species respond (clustering)

Within a covariate, there is homogeneity in the way species respond (clustering)

Within a covariate, there is homogeneity in the way species respond (clustering)

Multivariate abundance data

- What do we want?
  - Variable selection (species respond to a subset of covariates)
  - Homogeneity pursuit (group species according to their responses to each covariate)

- Ideally, the statistical method also:
  - Accounts for between-species covariation (residual correlations between columns of Y)

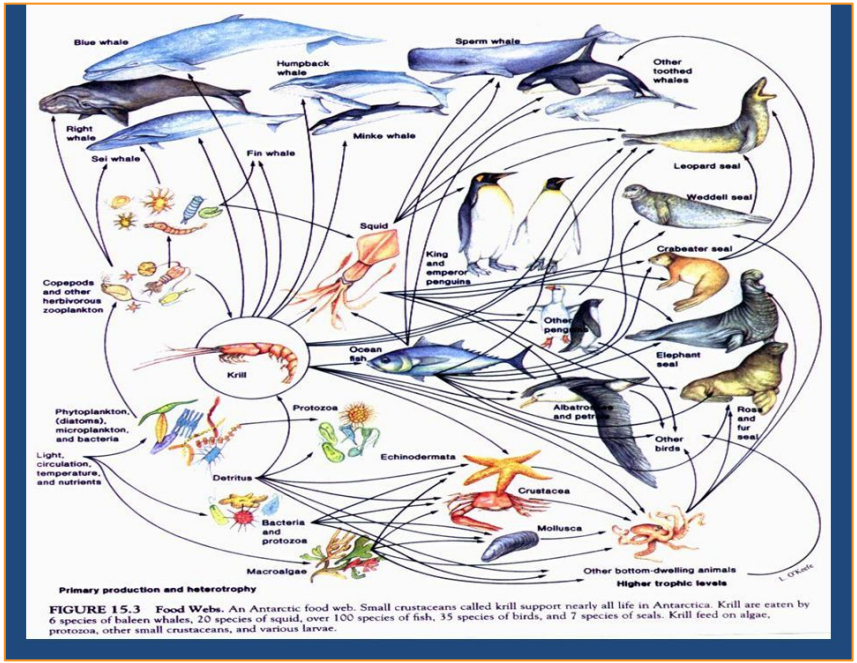


FIGURE 15.3 Food Webs. An Antarctic food web. Small crustaceans called krill support nearly all life in Antarctica. Krill are eaten by 6 species of baleen whales, 20 species of squid, over 100 species of fish, 35 species of birds, and 7 species of seals. Krill feed on algae, protozoa, other small crustaceans, and various larvae.

Generalized Estimating Equations

- GEEs for multivariate abundance data
  - Speedy-ish
  - Can account for residual correlations between species

Methods in Ecology and Evolution — BRITISH ECOLOGICAL SOCIETY

Free Access

mvabund– an R package for model-based analysis of multivariate abundance data

Yi Wang, Ulrike Naumann, Stephen T. Wright, David I. Warton

First published: 21 February 2012 | https://doi.org/10.1111/j.2041-210X.2012.00190.x | Citations: 328

Correspondence site: http://www.respond2articles.com/MEE/

ELSEVIER — Journal of Multivariate Analysis

Volume 143, January 2016, Pages 481-491

The analysis of multivariate longitudinal data using multivariate marginal models

Hyunkeun Cho

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.jmva.2015.10.012 Get rights and content

Under an Elsevier user license open archive

Journal of Computational and Graphical Statistics

Volume 27, 2018 - Issue 1

Submit an article Journal homepage

1,070 Views 12 CrossRef citations to date 3 Altmetric

Nonparametric

A Generalized Estimating Equation Approach to Multivariate Adaptive Regression Splines

Jakub Stoklosa & David I. Warton

Pages 245-253 | Received 01 Feb 2016, Published online: 06 Mar 2018

Cite this article https://doi.org/10.1080/10618600.2017.1360780

Full Article Figures & data References Supplemental Citations Metrics Reprints & Permissions

Read this article

Biometrics — JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY

Volume 67, March 2011, Pages 116-

Regularized Sandwich Estimators for Analysis of High-Dimensional Data Using Generalized Estimating Equations

David I. Warton

First published: 14 March 2011 | https://doi.org/10.1111/j.1541-0420.2010.01438.x | Citations: 68

email: David.Warton@unsw.edu.au

Read the full text PDF TOOLS SHARE

References Related

Recommended

Doubly robust generalized estimating equations for longitudinal data

Shaun Seaman, Andrew Co

Statistics in Medicine

PLOS COMPUTATIONAL BIOLOGY

BROWSE PUBLISH ABOUT SEARCH advanced search

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Generalized estimating equation modeling on correlated microbiome sequencing data with longitudinal measures

Bo Chen, Wei Xu

Version 2 Published: September 8, 2020 • https://doi.org/10.1371/journal.pcbi.1008108

36 Save 8 Citation 4,308 View 3 Share

Article Authors Metrics Comments Media Coverage Peer Review Download PDF

Generalized Estimating Equations

- GEEs for multivariate abundance data
  - Speedy-ish
  - Can account for residual correlations between species

Consider a set of N sites $\{(\boldsymbol{x}_i, \boldsymbol{y}_i); i = 1, \ldots, N\}$, where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iJ})^\top$, y_{ij} denotes the record for the j-th species at the i-th site, and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iP})^\top$ is a corresponding P-vector of covariates.

- Marginal mean: $g\{\mathsf{E}_Y(y_{ij})\} = g(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}_i^\top \boldsymbol{\beta}_j$,
- Marginal variance function: $\mathsf{Var}_Y(y_{ij}) = V(\mu_{ij}, \boldsymbol{\phi}_j)$ e.g., $V(\mu_{ij}, \boldsymbol{\phi}_j) = \mu_{ij}(1 - \mu_{ij})$ for binary responses;
- Marginal working covariance: $\mathsf{Cov}_Y(\boldsymbol{y}_i) = \boldsymbol{V}_i = \boldsymbol{A}_i^{1/2} \boldsymbol{R}(\boldsymbol{\gamma}) \boldsymbol{A}_i^{1/2}$, where $\boldsymbol{A}_i$ is a $J \times J$ diagonal matrix with entries $V(\mu_{ij}, \boldsymbol{\phi}_j)$

Generalized Estimating Equations

- GEEs for multivariate abundance data
  - Speedy-ish
  - Can account for residual correlations between species

Consider a set of N sites $\{(\boldsymbol{x}_i, \boldsymbol{y}_i); i = 1, \ldots, N\}$, where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iJ})^\top$, y_{ij} denotes the record for the j-th species at the i-th site, and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iP})^\top$ is a corresponding P-vector of covariates.

- Marginal mean: $g\{\mathsf{E}_Y(y_{ij})\} = g(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}_i^\top \boldsymbol{\beta}_j$,
- Marginal variance function: $\mathsf{Var}_Y(y_{ij}) = V(\mu_{ij}, \boldsymbol{\phi}_j)$ e.g., $V(\mu_{ij}, \boldsymbol{\phi}_j) = \mu_{ij}(1 - \mu_{ij})$ for binary responses;
- Marginal working covariance: $\mathsf{Cov}_Y(\boldsymbol{y}_i) = \boldsymbol{V}_i = \boldsymbol{A}_i^{1/2} \boldsymbol{R}(\boldsymbol{\gamma}) \boldsymbol{A}_i^{1/2}$, where $\boldsymbol{A}_i$ is a $J \times J$ diagonal matrix with entries $V(\mu_{ij}, \boldsymbol{\phi}_j)$

Solve

$$\boldsymbol{S}(\boldsymbol{B}) = \sum_{i=1}^{N} \boldsymbol{S}_i(\boldsymbol{B}) = \sum_{i=1}^{N} \boldsymbol{D}_i^\top \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}_{JP},$$

where $\boldsymbol{D}_i = \boldsymbol{W}_i \boldsymbol{X}_i$, $\boldsymbol{X}_i = \boldsymbol{I}_J \otimes \boldsymbol{x}_i^\top$ is a $J \times JP$ model matrix, $\boldsymbol{W}_i$ is a $J \times J$ diagonal matrix of weights, and

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{\beta}_1^\top \\ \boldsymbol{\beta}_2^\top \\ \vdots \\ \boldsymbol{\beta}_J^\top \end{bmatrix}$$

Generalized Estimating Equations

- GEEs for multivariate abundance data
  - Speedy-ish
  - Can account for residual correlations between species

- How to set up the working correlation?
  - Rank-reduced form

$$\boldsymbol{R}(\boldsymbol{\gamma}) = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top + \text{diag}(\xi, \ldots, \xi_J), \quad \text{where } \boldsymbol{\Gamma} \text{ is a } J \times L \text{ matrix. Pick } L \ll J$$

- How to estimate the (other) parameters?
  - Moment/quasi-likelihood estimation

Thinking about homogeneity pursuit

- Suppose we have 5 species

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{\beta}_1^\top \\ \boldsymbol{\beta}_2^\top \\ \vdots \\ \boldsymbol{\beta}_5^\top \end{bmatrix} = \begin{bmatrix} \beta_{11} & \dots & \beta_{1k} = -0.5 & \dots & \beta_{1P} \\ \beta_{21} & \dots & \beta_{2k} = 0.5 & \dots & \beta_{2P} \\ \beta_{31} & \dots & \beta_{3k} = 0.5 & \dots & \beta_{3P} \\ \beta_{41} & \dots & \beta_{4k} = -0.5 & \dots & \beta_{4P} \\ \beta_{51} & \dots & \beta_{5k} = 0.5 & \dots & \beta_{5P} \end{bmatrix}$$

Thinking about homogeneity pursuit

- Suppose we have 5 species

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{\beta}_1^\top \\ \boldsymbol{\beta}_2^\top \\ \vdots \\ \boldsymbol{\beta}_5^\top \end{bmatrix} = \begin{bmatrix} \beta_{11} & \dots & \beta_{1k} = -0.5 & \dots & \beta_{1P} \\ \beta_{21} & \dots & \beta_{2k} = 0.5 & \dots & \beta_{2P} \\ \beta_{31} & \dots & \beta_{3k} = 0.5 & \dots & \beta_{3P} \\ \beta_{41} & \dots & \beta_{4k} = -0.5 & \dots & \beta_{4P} \\ \beta_{51} & \dots & \beta_{5k} = 0.5 & \dots & \beta_{5P} \end{bmatrix}$$

- Write them in ascending order

$\beta_{(1),k} = \beta_{(2),k} = -0.5$, $\beta_{(3),k} = \beta_{(4),k} = \beta_{(5),k} = 0.5$. There are $J_{02} = 2 < J$ distinct elements.

Thinking about homogeneity pursuit

- Suppose we have 5 species

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{\beta}_1^\top \\ \boldsymbol{\beta}_2^\top \\ \vdots \\ \boldsymbol{\beta}_5^\top \end{bmatrix} = \begin{bmatrix} \beta_{11} & \dots & \beta_{1k} = -0.5 & \dots & \beta_{1P} \\ \beta_{21} & \dots & \beta_{2k} = 0.5 & \dots & \beta_{2P} \\ \beta_{31} & \dots & \beta_{3k} = 0.5 & \dots & \beta_{3P} \\ \beta_{41} & \dots & \beta_{4k} = -0.5 & \dots & \beta_{4P} \\ \beta_{51} & \dots & \beta_{5k} = 0.5 & \dots & \beta_{5P} \end{bmatrix}$$

- Write them in ascending order

$\beta_{(1),k} = \beta_{(2),k} = -0.5$, $\beta_{(3),k} = \beta_{(4),k} = \beta_{(5),k} = 0.5$. There are $J_{02} = 2 < J$ distinct elements.

- Consider the ordered successive differences

$$\begin{aligned} \beta_{(2),k} - \beta_{(1),k} &= 0 \\ \beta_{(3),k} - \beta_{(2),k} &= 1 \\ \beta_{(4),k} - \beta_{(3),k} &= 0 \\ \beta_{(5),k} - \beta_{(4),k} &= 0 \end{aligned}$$

- If we want homogeneity/clustering, then we want to shrink ordered successive differences to zero

Thinking about sparsity

- Suppose we have 5 species

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{\beta}_1^\top \\ \boldsymbol{\beta}_2^\top \\ \vdots \\ \boldsymbol{\beta}_5^\top \end{bmatrix} = \begin{bmatrix} \beta_{11} & \dots & \beta_{1k} = -0.25 & \dots & \beta_{1P} \\ \beta_{21} & \dots & \beta_{2k} = -0.5 & \dots & \beta_{2P} \\ \beta_{31} & \dots & \beta_{3k} = 0 & \dots & \beta_{3P} \\ \beta_{41} & \dots & \beta_{4k} = -0.5 & \dots & \beta_{4P} \\ \beta_{51} & \dots & \beta_{5k} = 0 & \dots & \beta_{5P} \end{bmatrix}$$

Thinking about sparsity

- Suppose we have 5 species

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{\beta}_1^\top \\ \boldsymbol{\beta}_2^\top \\ \vdots \\ \boldsymbol{\beta}_5^\top \end{bmatrix} = \begin{bmatrix} \beta_{11} & \dots & \beta_{1k} = -0.25 & \dots & \beta_{1P} \\ \beta_{21} & \dots & \beta_{2k} = -0.5 & \dots & \beta_{2P} \\ \beta_{31} & \dots & \beta_{3k} = 0 & \dots & \beta_{3P} \\ \beta_{41} & \dots & \beta_{4k} = -0.5 & \dots & \beta_{4P} \\ \beta_{51} & \dots & \beta_{5k} = 0 & \dots & \beta_{5P} \end{bmatrix}$$

- Write them in ascending order by their absolute value

$$\beta_{(|1|),k} = \beta_{(|2|),k} = 0,\ \beta_{(|3|),k} = -0.25,\ \beta_{(|4|),k} = \beta_{(|5|),k} = -0.5$$

- If there is sparsity, the coefficient with the smallest absolute value must be zero

Thinking about homogeneity pursuit + sparsity

- Key points. For each covariate:
  a. To group species into a smaller number of “canonical” coefficients, shrink ordered successive differences to zero
  b. To achieve sparsity, shrink the coefficient with the smallest absolute value to zero.
     - Note *only* the smallest absolute value coefficient is needed!

Thinking about homogeneity pursuit + sparsity

- Key points. For each covariate:
  a. To group species into a smaller number of "canonical" coefficients, shrink ordered successive differences to zero
  b. To achieve sparsity, shrink the coefficient with the smallest absolute value to zero.
     - Note *only* the smallest absolute value coefficient is needed!

- Augment the GEE with a penalty e.g., something based on

$$\mathcal{P}_\lambda = \lambda \sum_{k=2}^{P} \left(w_{1k} |\beta_{(|1|),k}| + \sum_{j=2}^{J} w_{jk} |\beta_{(j),k} - \beta_{(j-1),k}| \right)$$

Adaptive lasso of the smallest absolute value

Adaptive fused lasso of the ordered successive differences

HPGEE

- Penalized GEE: Solve

$$\boldsymbol{S}_{\text{pen}}(\boldsymbol{B}) = \sum_{i=1}^{N} \boldsymbol{D}_i^\top \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) - \frac{d\mathcal{P}_\lambda}{d\boldsymbol{B}} = \boldsymbol{0}_{JP}, \text{ where}$$

$$\mathcal{P}_\lambda = \lambda \sum_{k=2}^{P} \left(w_{1k}|\beta_{(|1|),k}| + \sum_{j=2}^{J} w_{jk}|\beta_{(j),k} - \beta_{(j-1),k}| \right)$$

- On the surface, this looks pretty challenging!

HPGEE (miracle in progress…)

- Turns out this is not too hard through reparametrization

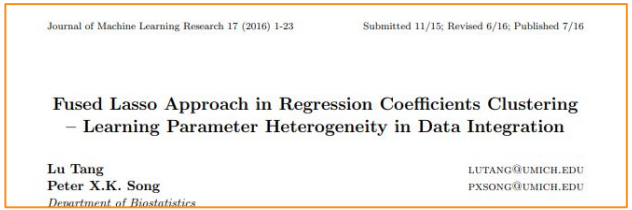
Define

$$\begin{bmatrix} \beta_{(|1|),1} & \cdots & \beta_{(|1|),k} & \cdots \beta_{(|1|),P} \\ \beta_{(2),1} - \beta_{(1),1} & \cdots & \beta_{(2),k} - \beta_{(1),k} & \cdots \beta_{(2),P} - \beta_{(1),P} \\ \beta_{(3),1} - \beta_{(2),1} & \cdots & \beta_{(3),k} - \beta_{(2),k} & \cdots \beta_{(3),P} - \beta_{(2),P} \\ \vdots & \ddots & \vdots & \ddots \quad \vdots \\ \beta_{(J),1} - \beta_{(J-1),1} & \cdots & \beta_{(J),k} - \beta_{(J-1),k} & \cdots \beta_{(J),P} - \beta_{(J-1),P} \end{bmatrix} = \begin{bmatrix} \upsilon_{11} & \cdots & \upsilon_{1k} & \cdots & \upsilon_{1P} \\ \upsilon_{21} & \cdots & \upsilon_{2k} & \cdots & \upsilon_{2P} \\ \upsilon_{31} & \cdots & \upsilon_{3k} & \cdots & \upsilon_{3P} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \upsilon_{J1} & \cdots & \upsilon_{Jk} & \cdots & \upsilon_{JP} \end{bmatrix} = \boldsymbol{\Upsilon}.$$

Then a vectorized version of marginal mean of the GEE can be written as

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{X}\text{vec}(\boldsymbol{B}^\top) = \check{\boldsymbol{X}}\text{vec}(\boldsymbol{\Upsilon}^\top)$$

where $\text{vec}(\boldsymbol{\Upsilon}^\top) = \boldsymbol{M}\text{vec}(\boldsymbol{B}^\top)$ and $\check{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{M}^{-1}$, and $\boldsymbol{M}$ is a $JP \times JP$ sparse, invertible matrix whose elements are a function of $\boldsymbol{B}$.


Journal of Machine Learning Research 17 (2016) 1-23 Submitted 11/15; Revised 6/16; Published 7/16

Fused Lasso Approach in Regression Coefficients Clustering – Learning Parameter Heterogeneity in Data Integration

Lu Tang LUTANG@UMICH.EDU
Peter X.K. Song PXSONG@UMICH.EDU
Department of Biostatistics


Biometrics JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY

BIOMETRIC METHODOLOGY

Poststratification fusion learning in longitudinal data analysis

Lu Tang Peter X.-K. Song

First published: 19 July 2020 | https://doi.org/10.1111/biom.13333

Volume 77, Issue 3
September 2021
Pages 914-928

References Related Information

Recommended

HPGEE (miracle in progress…)

- Rewrite the GEE…

  Redefine the marginal mean as $g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \check{\boldsymbol{X}}\text{vec}(\boldsymbol{\Upsilon}^\top)$, and write $\tilde{\boldsymbol{D}}_i = \boldsymbol{W}_i\check{\boldsymbol{X}}_i$. Then solve

$$\boldsymbol{S}(\boldsymbol{\Upsilon}) = \sum_{i=1}^{N} \boldsymbol{S}_i(\boldsymbol{\Upsilon}) = \sum_{i=1}^{N} \tilde{\boldsymbol{D}}_i^\top \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}_{JP}.$$

- ….and the penalty

$$\mathcal{P}_\lambda = \lambda \sum_{k=2}^{P} \sum_{j=1}^{J} w_{jk} |\upsilon_{j,k}|$$

HPGEE (a few details)

Definition 1. Given tuning parameter $\lambda > 0$, solve the HPGEE

$$\boldsymbol{S}_{\text{HPGEE}}(\boldsymbol{\Upsilon}) = \boldsymbol{S}(\boldsymbol{\Upsilon}) - \frac{d\mathcal{P}_\lambda}{d\boldsymbol{\Upsilon}} = \mathbf{0}_{JP}, \text{ where } \mathcal{P}_\lambda = \lambda \sum_{k=2}^{P} \sum_{j=1}^{J} w_{jk} |\upsilon_{jk}|$$

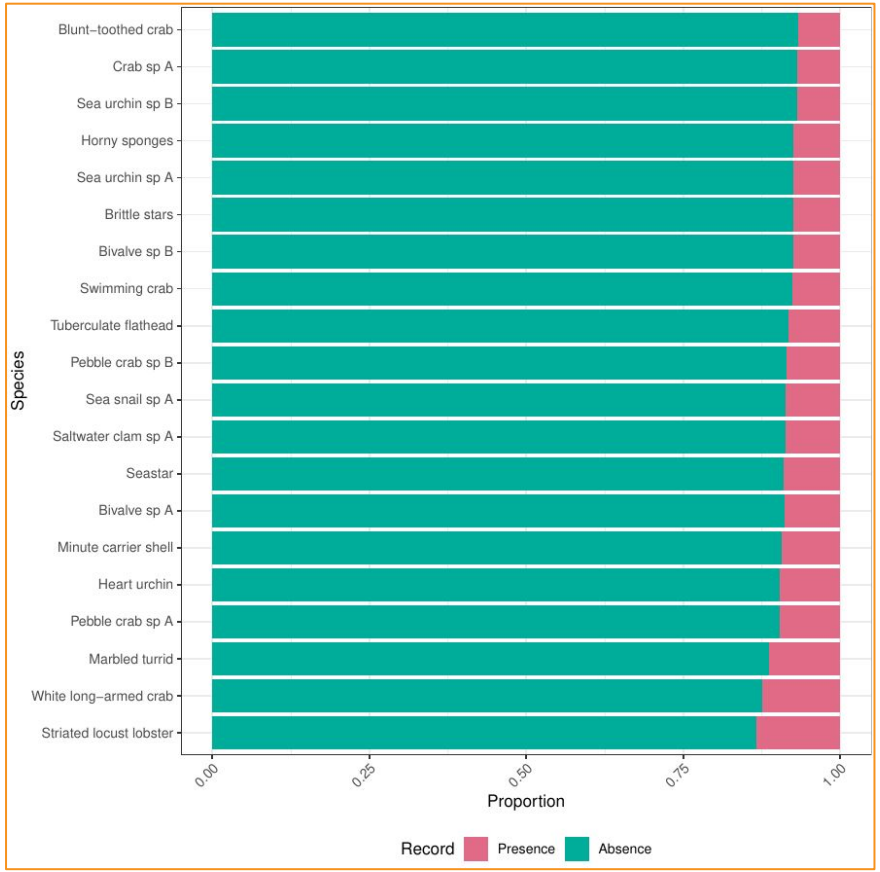
and w_{jk} are set of adaptive weights constructed from the unpenalized GEE.

- Computationally, the problem is much easier

Adaptive lasso of the smallest absolute value + Adaptive fused lasso of the ordered successive differences → Adaptive lasso on reparametrized coefficients

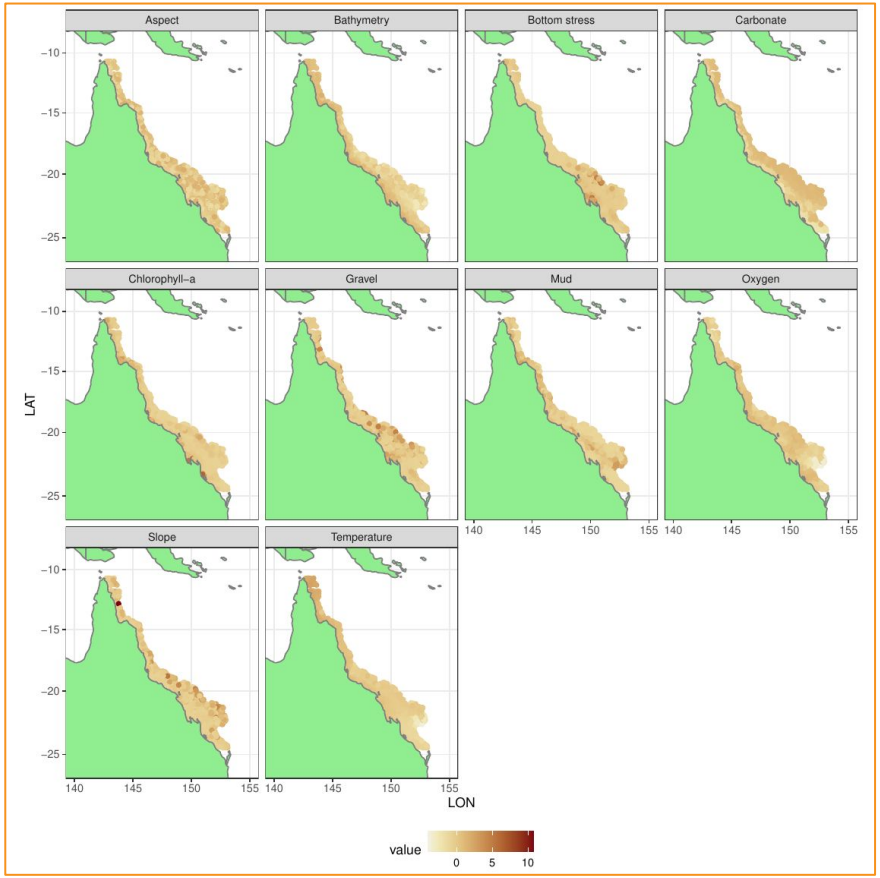
Application to Great Barrier Reef biodiversity

- Presence-absence records
  - J = 20 species; N = 1146 sites



Application to Great Barrier Reef biodiversity

- Presence-absence records
  - J = 20 species; N = 1146 sites
- Ten environmental predictors + intercept
  - P = 11 covariates



Application to Great Barrier Reef biodiversity

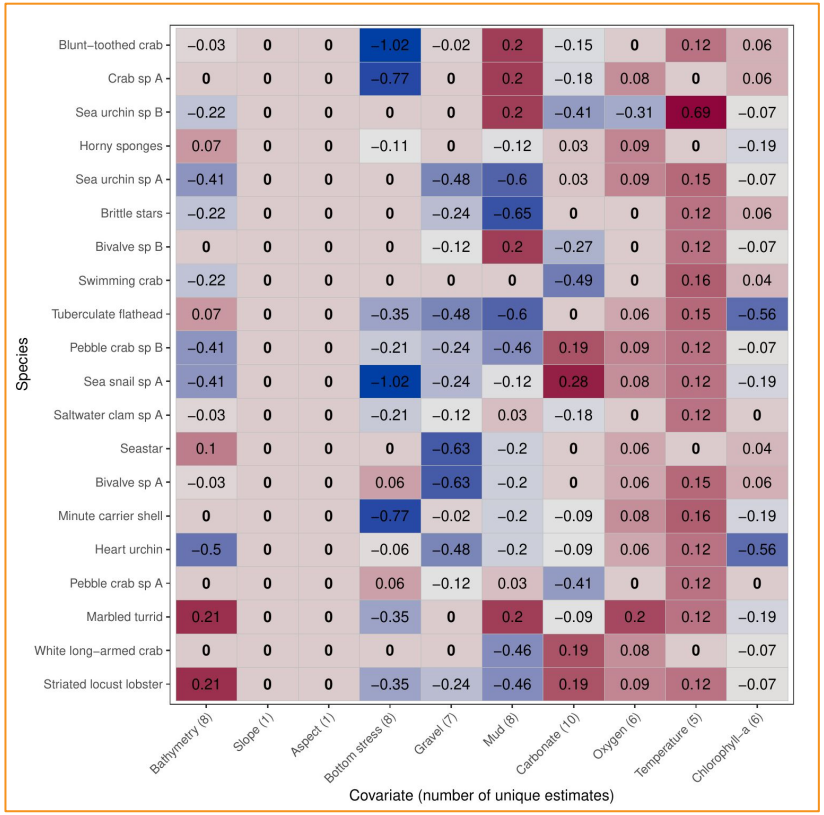
- Presence-absence records
  - J = 20 species; N = 1146 sites
- Ten environmental predictors + intercept
  - P = 11 covariates

- Apply HPGEE with all covariates as linear terms

- Marginal mean: $\Phi^{-1}(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}_i^\top \boldsymbol{\beta}_j$;
- Marginal variance function: $\mathrm{Var}_Y(y_{ij}) = \mu_{ij}(1 - \mu_{ij})$;
- Marginal working covariance: $\mathrm{Cov}_Y(\boldsymbol{y}_i) = \boldsymbol{V}_i = \boldsymbol{A}_i^{1/2} \boldsymbol{R}(\boldsymbol{\gamma}) \boldsymbol{A}_i^{1/2}$, where $\boldsymbol{R}(\boldsymbol{\gamma})$ has a rank-reduced form with rank $L = 3$.

Application to Great Barrier Reef biodiversity

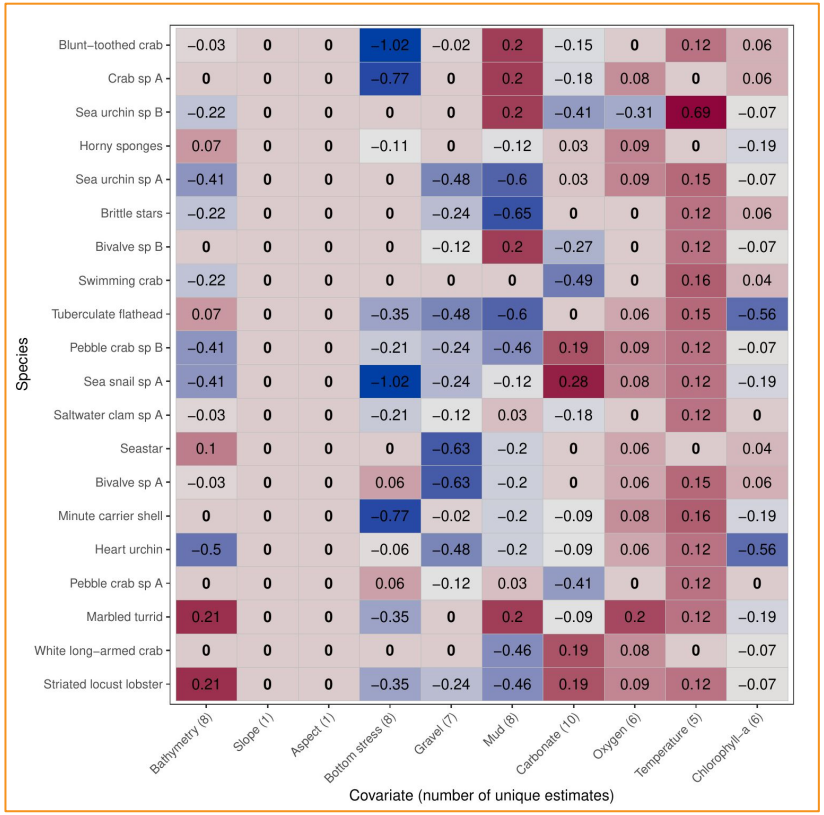
- Some level of sparsity
  - Slope and aspect non-informative for all species;
  - Percent mud and Chlorophyll-a most informative



Application to Great Barrier Reef biodiversity

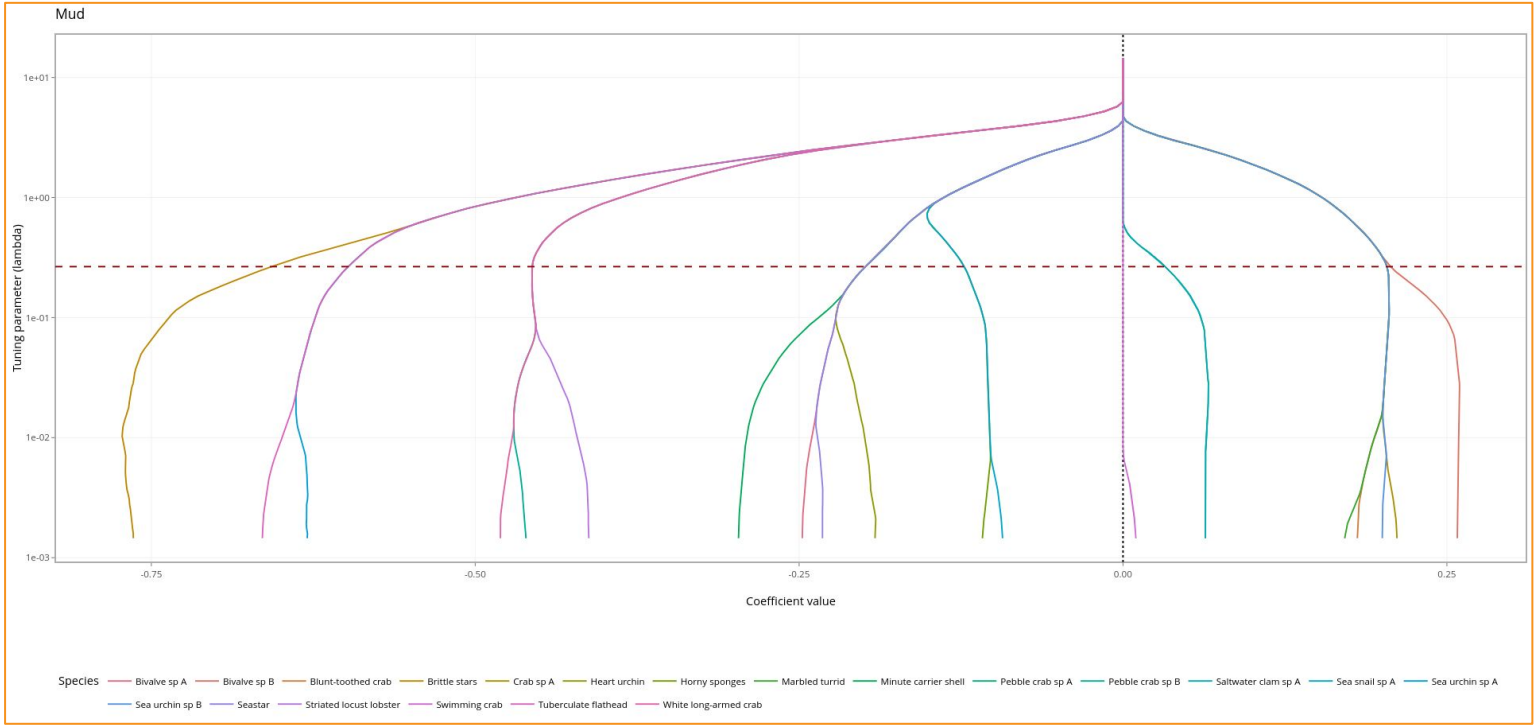
- Some level of sparsity
  - Slope and aspect non-informative for all species;
  - Percent mud and Chlorophyll-a most informative

- Lots of homogeneity
  - 200 individual slopes compressed to 60 canonical coefficients
  - Mean annual temperature is grouped into five non-negative canonical coefficients



Application to Great Barrier Reef biodiversity

- Sparsity + homogeneity: Example with percent mud



Closing remarks

- Manuscript accepted in *Biometrics*
  - Simulation study
  - Assess predictive performance in GBR application

- https://github.com/fhui28/HPGEE

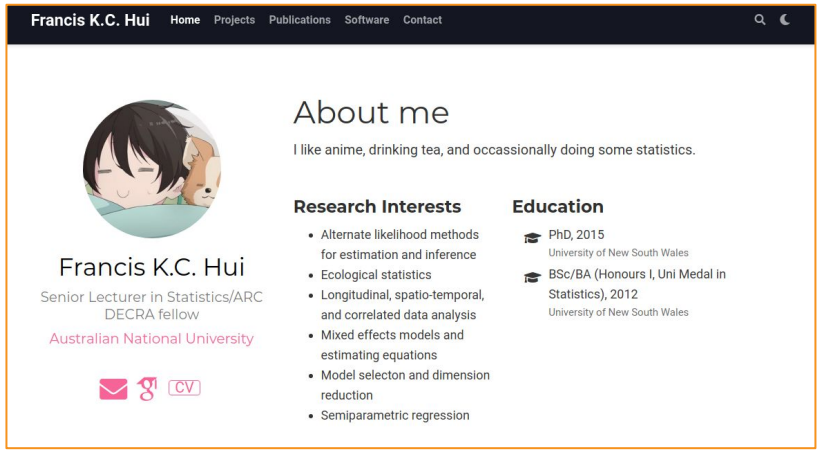
- HPGEE != Species Archetype Model/Species guilds
  - Clustering of species within covariates as opposed to entire their environmental response (parsimony versus flexibility)

- Countless extensions e.g., spatial/temporal correlations, more flexible regression models, large sample theory etc...

Thanks for listening!

Questions?

- francis.hui@anu.edu.au
- https://francishui.netlify.app/



An alternative to spatio-temporal LVMs that is more scalable

(but hopefully about as flexible?)

Generalized Estimating Equations

- GEEs for multivariate abundance data
  - Speedy-ish
  - Can account for residual correlations between species

- How to set up the working correlation?
  - Rank-reduced form

    $\boldsymbol{R}(\boldsymbol{\gamma}) = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top + \text{diag}(\xi, \ldots, \xi_J)$, where $\boldsymbol{\Gamma}$ is a $J \times L$ matrix. Pick $L \ll J$
  - Can also used other forms like independence and rely on the robustness property of GEE (but lose efficiency)

- How to estimate the (other) parameters?
  - Moment/quasi-likelihood estimators; discuss later

HPGEE

- Fit the unpenalized GEE

Solve $\boldsymbol{S}(\boldsymbol{B}) = \sum_{i=1}^{N} \boldsymbol{S}_i(\boldsymbol{B}) = \sum_{i=1}^{N} \boldsymbol{D}_i^\top \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}_{JP}$ and construct an $\boldsymbol{M}$ from this.

HPGEE

- Fit the unpenalized GEE

Solve $\boldsymbol{S}(\boldsymbol{B}) = \sum_{i=1}^{N} \boldsymbol{S}_i(\boldsymbol{B}) = \sum_{i=1}^{N} \boldsymbol{D}_i^\top \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}_{JP}$ and construct an $\boldsymbol{M}$ from this.

- Homogeneity pursuit and variable selection in GEEs

Definition 1. For a given tuning parameter $\lambda > 0$, solve the HPGEE

$$\boldsymbol{S}_{\text{HPGEE}}(\boldsymbol{\Upsilon}) = \boldsymbol{S}(\boldsymbol{\Upsilon}) - \lambda \sum_{k=2}^{P} \sum_{j=1}^{J} w_{jk} \text{sgn}(\upsilon_{jk}) = \boldsymbol{0}_{JP},$$

where the w_{jk}'s are set of adaptive weights (also) constructed from the unpenalized GEE.

  - Note species-specific intercept not penalized
  - Please see the paper for details regarding constructing of adaptive weights

HPGEE (a few details)

Definition 1. Given tuning parameter $\lambda > 0$, solve the HPGEE

$$\boldsymbol{S}_{\text{HPGEE}}(\boldsymbol{\Upsilon}) = \boldsymbol{S}(\boldsymbol{\Upsilon}) - \frac{d\mathcal{P}_\lambda}{d\boldsymbol{\Upsilon}} = \mathbf{0}_{JP}, \text{ where } \mathcal{P}_\lambda = \lambda \sum_{k=2}^{P} \sum_{j=1}^{J} w_{jk} |\upsilon_{jk}|$$

and w_{jk} are set of adaptive weights constructed from the unpenalized GEE.

- Computationally, the problem is much easier

Adaptive lasso of the smallest absolute value + Adaptive fused lasso of the ordered successive differences → Adaptive lasso on reparametrized coefficients

  - Iteratively solve a penalized generalized least squares problem e.g., `glmnet`
  - Maximum pseudo-likelihood estimation to solve dispersion and working correlation matrix e.g., `factanal`

HPGEE (a few details)

Definition 1. For a given tuning parameter $\lambda > 0$, solve the HPGEE

$$\boldsymbol{S}_{\text{HPGEE}}(\boldsymbol{\Upsilon}) = \boldsymbol{S}(\boldsymbol{\Upsilon}) - \lambda \sum_{k=2}^{P} \sum_{j=1}^{J} w_{jk} \text{sgn}(\upsilon_{jk}) = \mathbf{0}_{JP},$$

where the w_{jk}'s are set of adaptive weights (also) constructed from the unpenalized GEE.

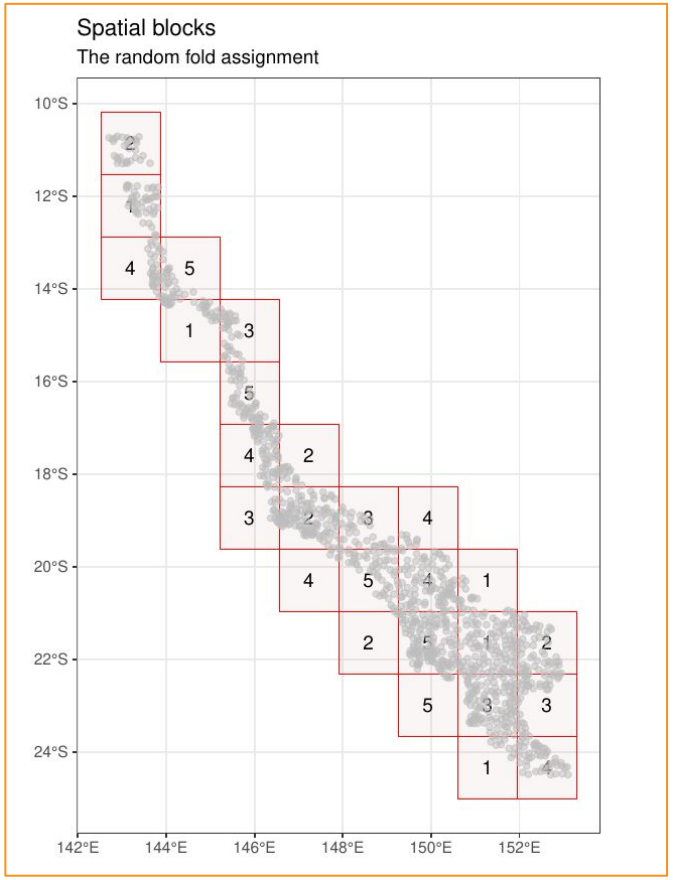
- Tuning parameter selection: Score Information Criterion

$$\text{SIC}_{\tau}(\lambda) = \sum_{i=1}^{N} \boldsymbol{S}_i(\hat{\boldsymbol{\Upsilon}}_{\lambda})^{\top} \boldsymbol{I}_i(\hat{\boldsymbol{\Upsilon}})^{-1} \boldsymbol{S}_i(\hat{\boldsymbol{\Upsilon}}_{\lambda}) + \tau \sum_{k=2}^{P} \sum_{j=1}^{J} I(\hat{\upsilon}_{\lambda,jk} \neq 0).$$

Application to Great Barrier Reef biodiversity

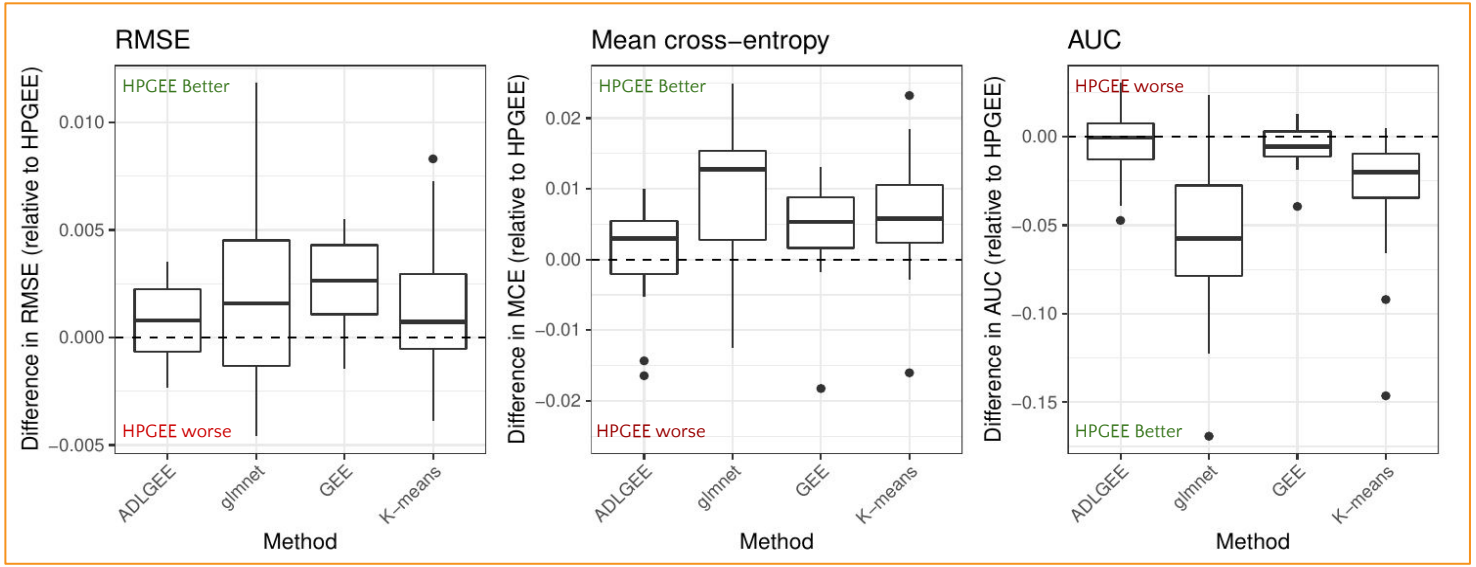
- Assess out-of-sample predictive performance
  - Five-fold block cross-validation (~80% training sites per fold) using `blockCV`

- Compare to four methods:
  - Penalized GEE with adaptive lasso (sparsity only)
  - `Glmnet` (sparsity only; independent species)
  - Unpenalized GEE (no sparsity or clustering)
  - GEE + K-Means (clustering only)



Application to Great Barrier Reef biodiversity

- Compare to four methods:
  - Penalized GEE with adaptive lasso (sparsity only)
  - `glmnet` (sparsity only; independent species)
  - Unpenalized GEE (no sparsity or clustering)
  - GEE + K-Means (clustering only)



Closing remarks

- Manuscript accepted in *Biometrics*
- https://github.com/fhui28/HPGEE
- HPGEE != Species Archetype Model/Species guilds
  - Clustering of species within covariates as opposed to entire their environmental response (parsimony versus flexibility)
- Can you do this for multivariate GLMMs, and joint species distribution/latent variable models?
  - Yes, but the computation becomes harder (work in progress)
- Countless extensions e.g., spatial/temporal correlations, more flexible regression models, large sample theory etc...