Copula inference for multivariate abundance data

Gordana Popovic, David Warton & Francis Hui

Eco-Stats Research Group, UNSW Sydney

December 2, 2015

Multivariate abundance data - Bush regeneration study

Multivariate abundance data - Bush regeneration study

Multivariate abundance data - Bush regeneration study

Anthony's data

Site	Treatment	Acarina	Blattodea	Collembola	...	Tricladida
1	0	21	3	1093	...	0
2	1	70	0	580	...	1
3	1	306	0	13541	...	0
4	1	98	0	2809	...	0
5	0	8	4	477	...	4
6	1	112	1	7527	...	0
...	...	...	...	...	...	...
10	1	320	0	5184	...	1

Data thanks to Anthony J. Pik at Macquarie University

Multivariate abundance data - Bush regeneration study

Anthony's data

Site	Treatment	Acarina	Blattodea	Collembola	...	Tricladida
1	0	21	3	1093	...	0
2	1	70	0	580	...	1
3	1	306	0	13541	...	0
4	1	98	0	2809	...	0
5	0	8	4	477	...	4
6	1	112	1	7527	...	0
...	...	...	...	...	...	...
10	1	320	0	5184	...	1

Data thanks to Anthony J. Pik at Macquarie University

Multivariate abundance data - Bush regeneration study

Anthony's data

Site	**Treatment**	Acarina	Blattodea	Collembola	...	Tricladida
1	0	21	3	1093	...	0
2	1	70	0	580	...	1
3	1	306	0	13541	...	0
4	1	98	0	2809	...	0
5	0	8	4	477	...	4
6	1	112	1	7527	...	0
...	...	...	...	...	...	...
10	1	320	0	5184	...	1

Data thanks to Anthony J. Pik at Macquarie University

Multivariate abundance data - Bush regeneration study

Anthony's data

Site	Treatment	Acarina	Blattodea	Collembola	...	Tricladida
1	0	21	3	1093	...	0
2	1	70	0	580	...	1
3	1	306	0	13541	...	0
4	1	98	0	2809	...	0
5	0	8	4	477	...	4
6	1	112	1	7527	...	0
...	...	...	...	...	...	...
10	1	320	0	5184	...	1

Data thanks to Anthony J. Pik at Macquarie University

Multivariate abundance data - Bush regeneration study

Anthony's data

Site	Treatment	Acarina	**Blattodea**	Collembola	**...**	**Tricladida**
1	0	21	3	1093	...	0
2	1	70	0	580	...	1
3	1	306	0	13541	...	0
4	1	98	0	2809	...	0
5	0	8	4	477	...	4
6	1	112	1	7527	...	0
...	...	...	...	...	...	...
10	1	320	0	5184	...	1

Data thanks to Anthony J. Pik at Macquarie University

Multivariate abundance data - Bush regeneration study

Anthony's data

Site	Treatment	**Acarina**	Blattodea	Collembola	...	Tricladida
1	0	21	3	1093	...	0
2	1	70	0	580	...	1
3	1	306	0	13541	...	0
4	1	98	0	2809	...	0
5	0	8	4	477	...	4
6	1	112	1	7527	...	0
...	...	...	...	...	...	...
10	1	320	0	5184	...	1

Data thanks to Anthony J. Pik at Macquarie University

Multivariate abundance data - Bush regeneration study

Anthony's data

Site	Treatment	Acarina	Blattodea	Collembola	...	Tricladida
1	0	21	3	1093	...	0
2	1	70	0	580	...	1
3	1	306	0	13541	...	0
4	1	98	0	2809	...	0
5	0	8	4	477	...	4
6	1	112	1	7527	...	0
...	...	...	...	...	...	...
10	1	320	0	5184	...	1

Question: Is there an effect of treatment?

Data thanks to Anthony J. Pik at Macquarie University

Outline

Inference for multivariate abundance data using copulas

- GEE inference for predictors, properties of Wald and Score
- Should we estimate dependence for inference?
- Building flexible multivariate models with copulas
- Simulations and ecological example

Outline

Inference for multivariate abundance data using copulas

- **GEE inference for predictors, properties of Wald and Score**
- Should we estimate dependence for inference?
- Building flexible multivariate models with copulas
- Simulations and ecological example

GEEs v.s. Likelihood based **inference** for predictors

- Generalised estimating equations (GEEs) are a procedure that fits models using score equations (Liang & Zeger 1986)
- GEEs fit models to correlated variables (e.g. Species) without specifying a multivariate model (likelihood)
- They can incorporate information about correlation between variables into parameter estimation and estimate correlation between model parameters
- We can use GEEs to carry out multivariate hypothesis testing with Wald and Score statistics
- Extensions can deal with data with small numbers of replicates (N) relative to the number of variables (P)

GEEs v.s. Likelihood based **inference** for predictors

	GEE	**Data**
Accommodate over-dispersion	✓	overdispersed
Accommodate large P small N	✓	P=24 N=10
Incorporate dependence	✓	species interact

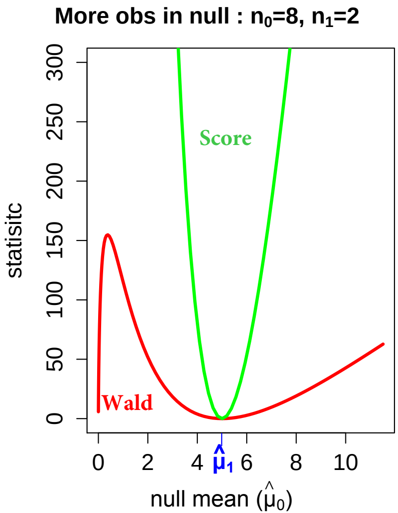
GEEs v.s. Likelihood based **inference** for predictors

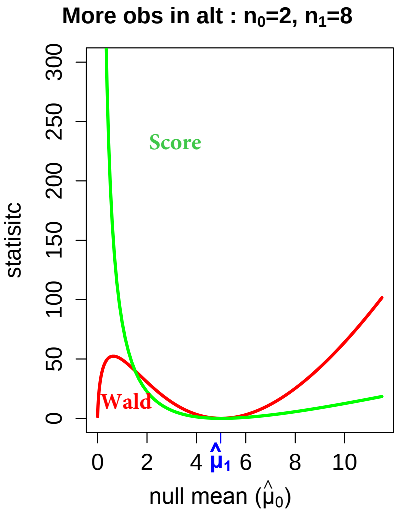
	GEE	**Data**
Accommodate over-dispersion	✓	overdispersed
Accommodate large P small N	✓	P=24 N=10
Incorporate dependence	✓	species interact
Likelihood based	X	

GEEs v.s. Likelihood based **inference** for predictors

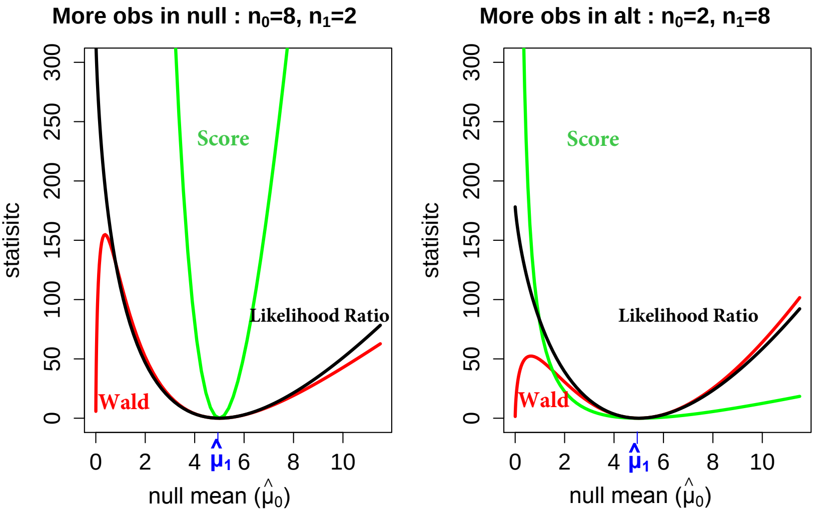
	GEE	**Data**
Accommodate over-dispersion	✓	overdispersed
Accommodate large P small N	✓	P=24 N=10
Incorporate dependence	✓	species interact
Likelihood based	X	
Good power for small means	not Wald	rare species
Good power in unbalanced designs	not Score	unbalanced

Wald and Score stat for unbalanced designs / small means





Wald and Score stat for unbalanced designs / small means

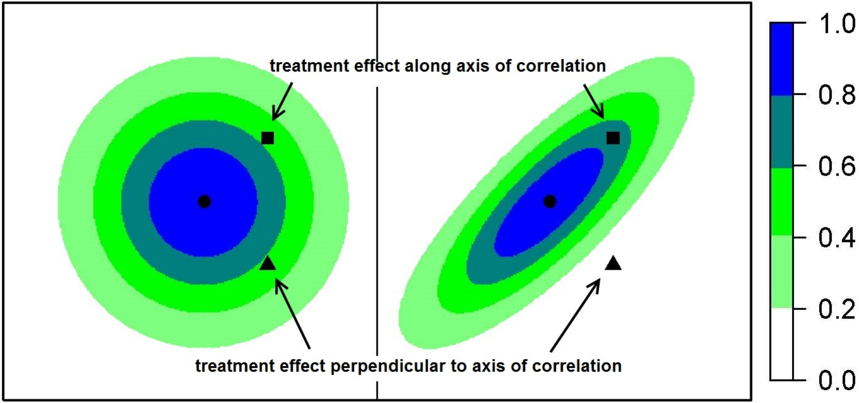


GEEs v.s. Likelihood based **inference** for predictors

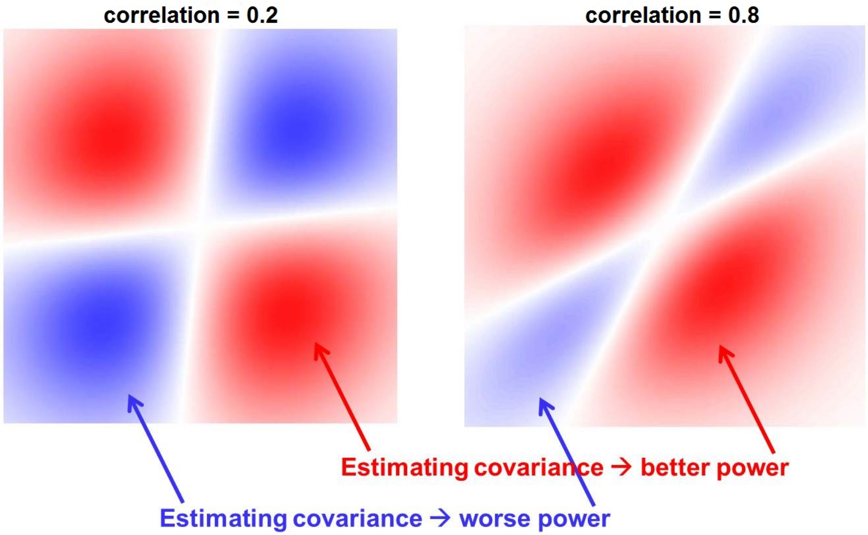
	GEE	**Data**
Accommodate over-dispersion	✓	overdispersed
Accommodate large P small N	✓	P=24 N=10
Incorporate dependence	✓	species interact
Likelihood based	X	
Good power for small means	not Wald	rare species
Good power in unbalanced designs	not Score	unbalanced
Likelihood ratio tests	only IID	

So do we want a method that can incorporate dependence AND uses likelihoods for inference?

Do we need to estimate dependence?



Do we need to estimate dependence?



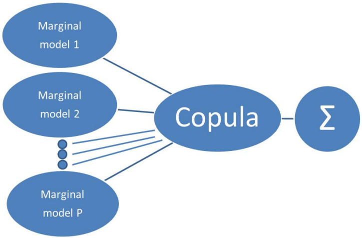
Want likelihood and dependence → Copulas

- Copulas stitch together **marginal distributions** and the **dependence structure** of a multivariate model. e.g

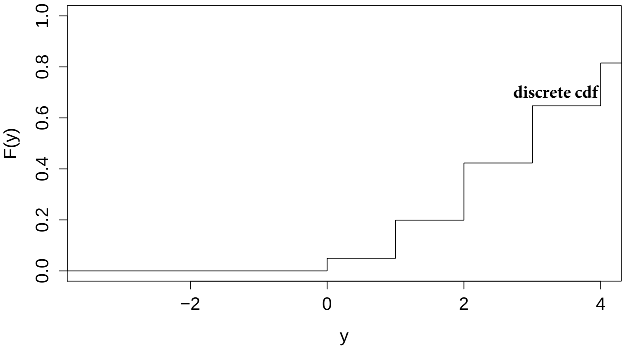
  Negative binomial marginals for (overdispersed) counts

  AND

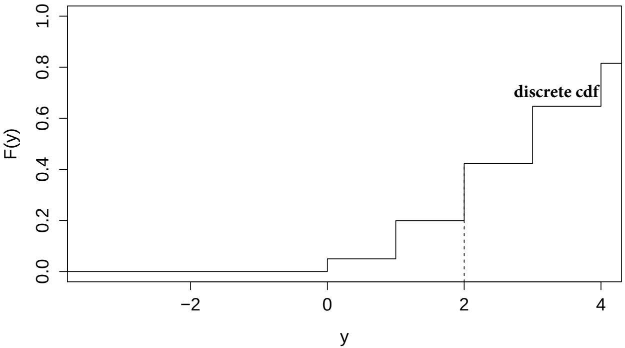
  The **dependence** structure of a **multivariate Normal**



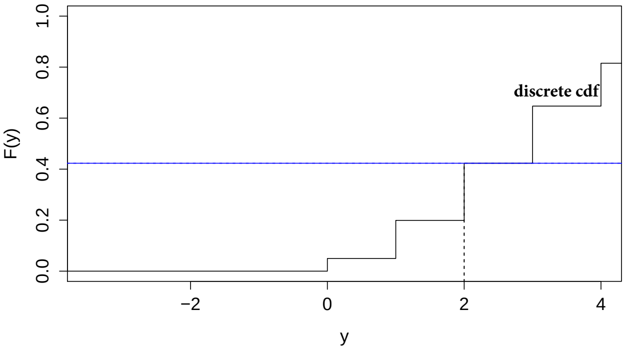
Gaussian copulas for discrete data



Gaussian copulas for discrete data

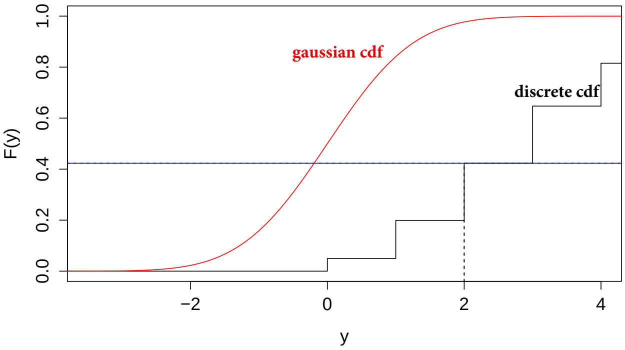


Gaussian copulas for discrete data



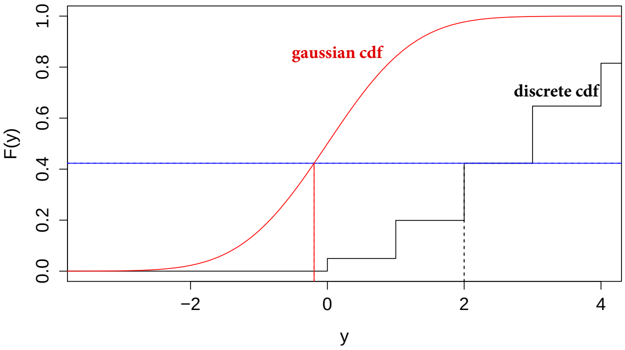
$$u_j = F(y_j)$$

Gaussian copulas for discrete data



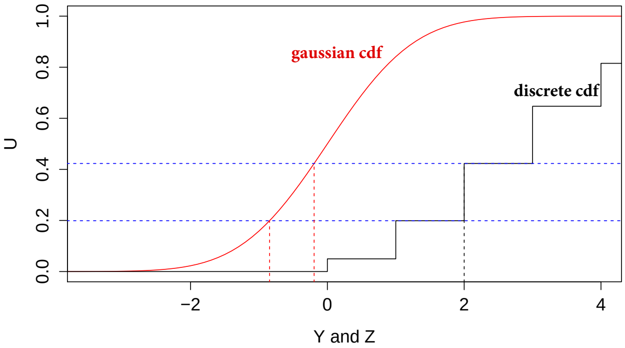
$$u_j = F(y_j)$$

Gaussian copulas for discrete data

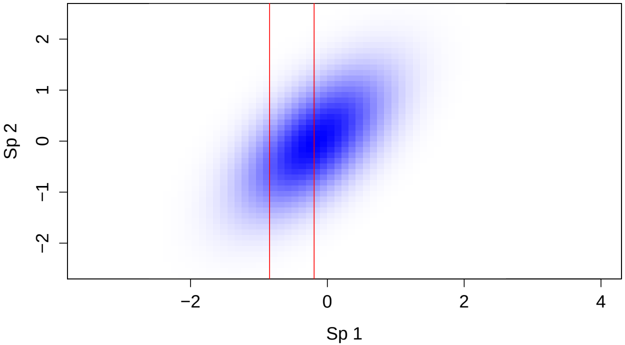


$$u_j = F(y_j) \qquad\qquad z = \Phi^{-1}(u)$$

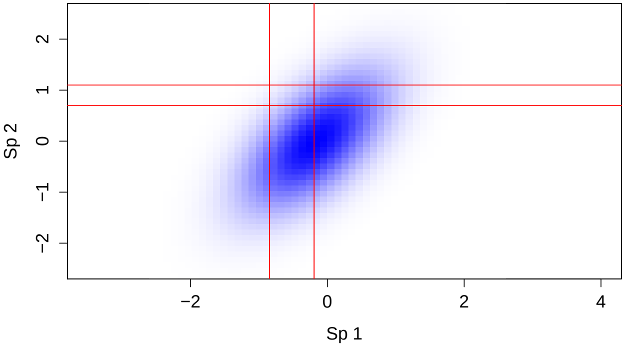
Gaussian copulas for discrete data



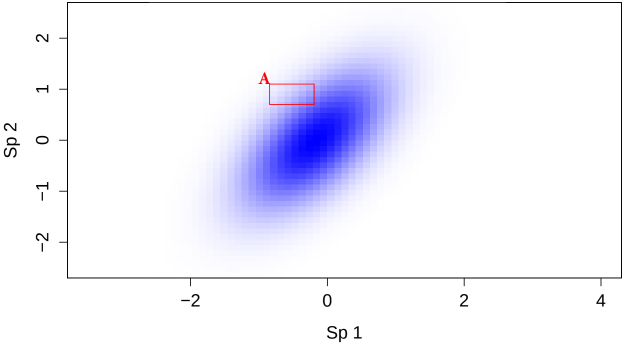
Gaussian copulas for discrete data



Gaussian copulas for discrete data

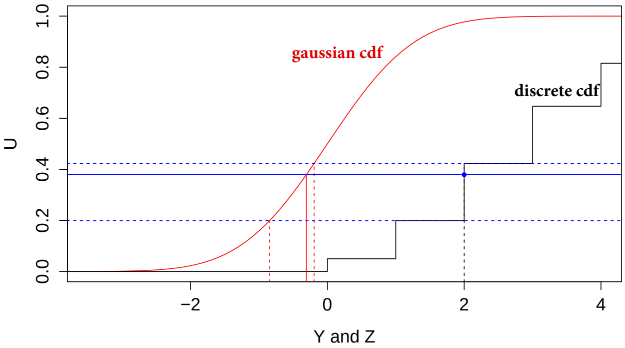


Gaussian copulas for discrete data



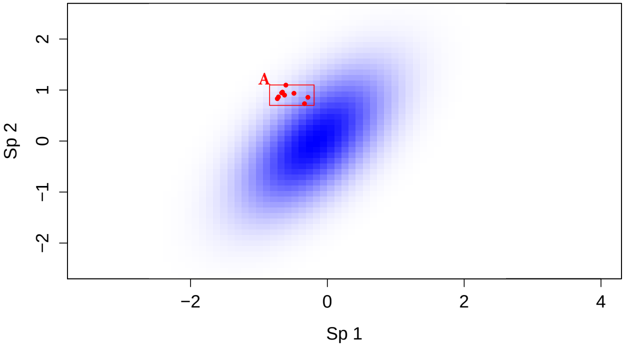
$$P(\mathbf{Y}_i = \mathbf{y}|\beta, \Sigma) = \int_A \cdots \int \phi(\mathbf{z}; \Sigma) d\mathbf{z}$$

Estimation with probability integral transform (PIT) resid



$$z_{ij} = \Phi^{-1}\left\{F_{ij}(y_{ij} - 1) + u_{ij} f_{ij}(y_{ij})\right\}$$

Estimation with probability integral transform (PIT) redid



$$\log L(\mathbf{y};\beta,\Sigma_\theta) \approx \left[\sum_i \sum_j \log(f_{ij}(y_{is},\beta_j))\right] + \sum_i \log\left[\sum_k \frac{\phi(\mathbf{z}_i^k;\Sigma_\theta)}{\prod_j \phi(z_{is}^k)}\right]$$

Copula likelihood

Copula Likelihood

$$L(\mathbf{Y} = \mathbf{y}|\beta, \Sigma) = \prod_i \int_A \cdots \int \phi(\mathbf{z}_i; \Sigma) d\mathbf{z}_i$$

Approximation by importance sampling

$$\log L(\mathbf{y}; \beta, \Sigma_\theta) \approx \left[\sum_i \sum_j \log(f_{ij}(y_{is}, \beta_j)) \right] + \sum_i \log \left[\sum_k \frac{\phi(\mathbf{z}_i^k; \Sigma_\theta)}{\prod_j \phi(z_{is}^k)} \right]$$

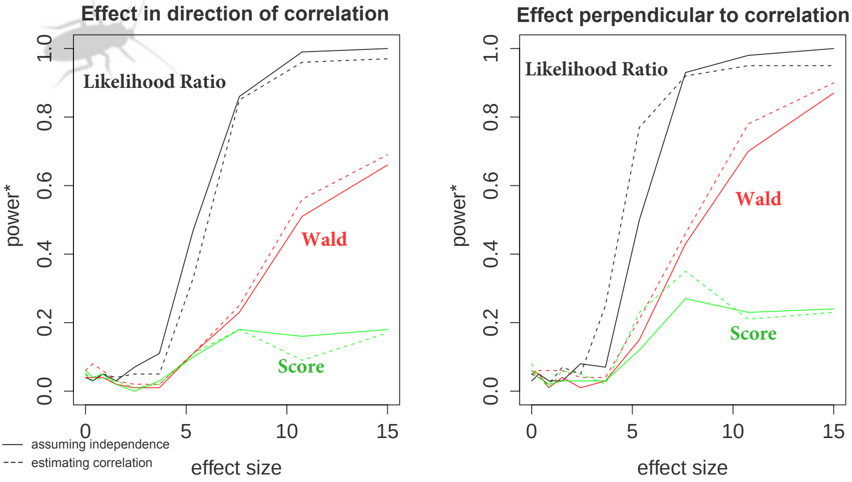
Estimate Σ using covariance modelling (Popovic et al., in Review)

- Unstructured
- Factor analysis
- Graphical model

GEEs v.s. Likelihood based **inference** for predictors

	GEE	**Data**
Accommodate over-dispersion	✓	overdispersed
Accommodate large P small N	✓	P=24 N=10
Incorporate dependence	✓	species interact
Likelihood based	X	
Good power for small means	not Wald	rare species
Good power in unbalanced designs	not Score	unbalanced
Likelihood ratio tests	only IID	

Simulation study for bush regeneration data



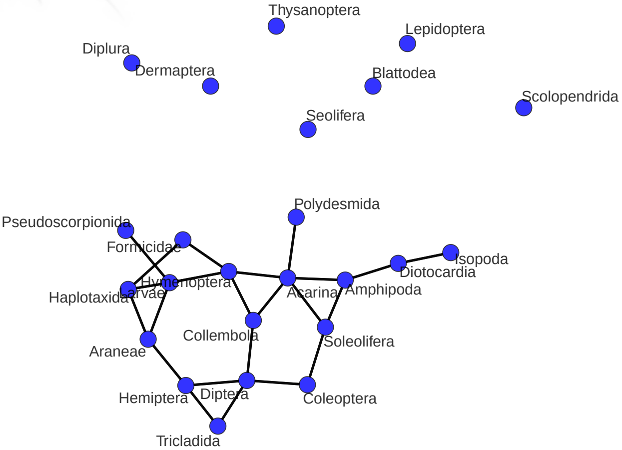
* Based on permutation of PIT residuals

Data Analysis - Test for effect of bush regeneration

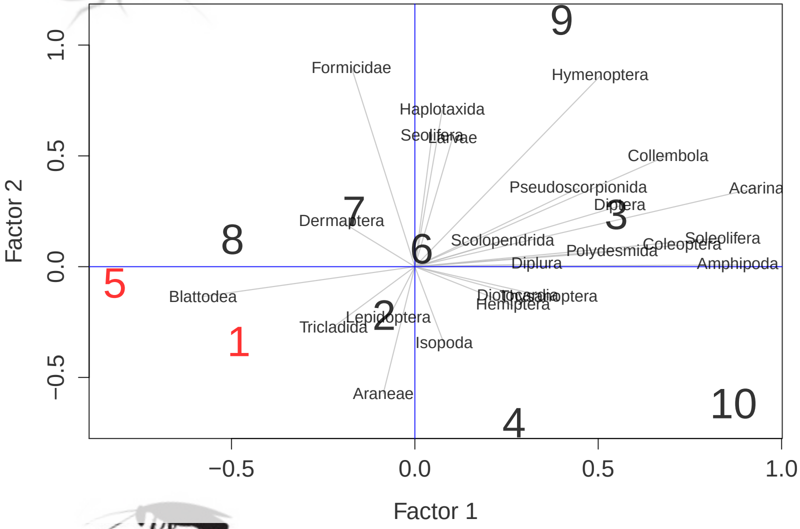
Method	P - value*
GEE Wald independent	0.031
GEE Score independent	0.244
Likelihood ratio test Independent	0.035
GEE Wald with dependence	0.028
GEE Score with dependence	0.307
Copula Likelihood ratio test with dependence	0.026

* Based on permutation of PIT residuals

Data Analysis - Graphical model (after accounting for trt)



Data Analysis - Biplot

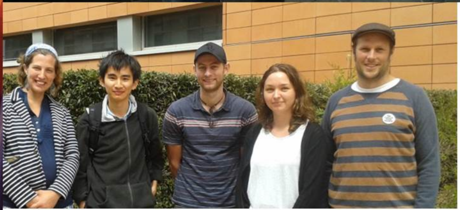


eco-stats

UNSW Ecological Statistics Research

MATHEMATICS & STATISTICS

THE UNIVERSITY OF NEW SOUTH WALES



Contact : g.popovic@unsw.edu.au

Data thanks to Anthony J. Pik at Macquarie University