Covariance modelling and inference for multivariate discrete data in ecology

Gordana Popovic

A thesis in fulfilment of the requirements for the degree of

Doctor of Philosophy

School of Mathematics and Statistics

Faculty of Science

September 2017

PLEASE TYPE

THE UNIVERSITY OF NEW SOUTH WALES
Thesis/Dissertation Sheet

Surname or Family name: **Popovic**

First name: **Gordana** Other name/s:

Abbreviation for degree as given in the University calendar: **PhD**

School: **School of Mathematics and Statistics** Faculty: **Faculty of Science**

Title: **Covariance modelling and inference for multivariate discrete data in ecology.**

Abstract 350 words maximum: (PLEASE TYPE)

In this thesis we use discrete copulas to develop novel methods to model multivariate abundance data in ecology. These data, which consist of measures of abundance for many species at a set of sites, occur naturally in ecological sampling. Multivariate abundance data are therefore very common, but also challenging to analyse. The responses are discrete and sparse, with many variables relative to sample size. We propose the use of Gaussian copulas, combined with covariance modelling, to create flexible models for multivariate abundance data that take the aforementioned properties into account. Copulas are not commonly used in ecology, but are well suited to modelling multivariate abundance data due to their flexibility. The modelling framework we propose extends the flexibility of copulas further, by combining any set of discrete or continuous response distributions, with any covariance modelling algorithm designed for Gaussian data. We first propose a novel estimation method for such models, and explore the use of these models to study patterns in correlation between species, using covariance modelling techniques. Then we introduce a tool to visualise species interactions using copula Gaussian graphical models. We demonstrate this on a large dataset of New Zealand native forest species, where we are able to uncover known species relationships as well as generate new hypotheses for how species interact. We then use Gaussian copula models to carry out marginal inference. In particular, when it comes to marginal hypothesis testing and model selection, the likelihood based inference implemented with Gaussian copulas has several advantages over the commonly used approach based on generalised estimating equations.

Declaration relating to disposition of project thesis/dissertation

I hereby grant to the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation in whole or in part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all property rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

I also authorise University Microfilms to use the 350 word abstract of my thesis in Dissertation Abstracts International (this is applicable to doctorat theses only).

.....

Sign

The University recognises that there may be exceptional circumstances requiring restrictions on copying or conditions on use. Requests for restriction for a period of up to 2 years must be made in writing. Requests for a longer period of restriction may be considered in exceptional circumstances and require the approval of the Dean of Graduate Research.

FOR OFFICE USE ONLY Date of completion of requirements for Award:

ORIGINALITY STATEMENT

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.'

Signed ……..

Date ……..

Acknowledgements

I'd like to acknowledge my family and friends, who have supported me throughout this PhD. To my parents, thank you for encouraging my curiosity, and for bringing me to a country with beautiful beaches, where a person can spend their whole life learning and not go hungry.

To Natty, who gave me perspective when I needed it the most, thank you for making laugh. My sister (from another mister) Carleigh, who took me in and listened when I needed to rant and always encouraged me. To Úna, thank you for making me forget my stresses whenever I was with you.

All the friends that have supported me along the way; Ben, who never let me wallow in misery when I could be climbing; Brenton, who was always ready to take me in and philosophise; Susannah and Megan, who kept me sane with coffee and friendship; and Sam, Firouzeh, Thomas, Daniel, and the many more who gave me their friendship and love.

I want to sincerely thank my supervisors, David and Francis. I could not have asked for better supervisors. They are always encouraging, knowledgeable, and fun, and without them I could not have finished this PhD.

And lastly, caffeine, to whom I dedicate this PhD. You have been with me from day one, and without you I am nothing.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this thesis we develop new methods for analysing multivariate abundance data in ecology. These data are routinely collected to study how community structure changes in response to the environment. They consist of measurements of abundance of plants or animals (most commonly presence/absence, counts, cover or biomass) simultaneously collected for a large number of taxa, with the intention of making inference about the community as a whole, rather than individual species or taxa. In Section 1.1 we describe three such datasets, which we will use as motivating examples in this thesis. These data are collected to study how environmental factors are associated with the distribution of species, as well as how species interact with one another.

From the point of view of analysis, multivariate abundance data present several challenges. They are highly discrete, with a large portion of observed absences, as many taxa will only be observed at a few sites. Sample sizes are often small, particularly when considered relative to the dimension of the response, with the number of response variables generally of the same order and sometimes exceeding the sample size (Table 1.1).

The main contribution of this thesis is a flexible and powerful method for modelling multivariate abundance data with Gaussian copulas. We begin in Chapter 2 by reviewing how multivariate abundance data are currently being modelled. Methods include generalised estimating equations for marginal inference (Section 2.2) and

hierarchical models for conditional inference (Section 2.3). In Chapter 3 we review copulas. They are a flexible method for marginal multivariate inference, commonly used in finance (Cherubini et al., 2004) and engineering (Genest & Favre, 2007), but largely unexplored for ecology. They are well suited to modelling complex data, given their flexibility (copulas allow specification of marginal distributions and covariance structure independently of one another) and relative ease of estimation with maximum likelihood (Section 3).

The two dominant properties of multivariate abundance data are their discreteness, and their high dimension. Estimating the covariance between variables is challenging for these data, as there are a large number of response variables (taxa) compared to the sample size (number of sites). As a result, most models currently used either assume independence (Section 2.1), which has consequences for the power of inference (Section 7.1.1), or use latent variables (Section 2.3) to specify a parsimonious covariance matrix. Latent variables additionally provide information about the patterns of correlation between taxa. Latent variable modelling is one of several types of covariance modelling (Chapter 4). There are other covariance modelling paradigms, like graphical models (Section 4.1.2), which can elicit quite different information about correlations, but currently no method which allows these to be implemented for multivariate abundance data.

In Chapter 5 we propose a novel method for applying any covariance modelling algorithm intended for Gaussian data, to discrete data. This algorithm allows us to fit parsimonious multivariate models, which respect the key properties of multivariate abundance data. We can additionally investigate patterns in covariance between species with a variety of covariance models, and answer questions about how species interact. In Chapter 6 we demonstrate the power of these models by carrying out graphical modelling (Section 4.1.2) on a large and complex multivariate ordinal dataset. We demonstrate how Gaussian copula graphical models allow us to investigate species interactions in the presence of covariates, and can provide new insights into relationships between species.

The other key property of multivariate abundance data is their discreteness. The mean-variance relationship induced by the discreteness of the data is often modelled

with extensions of generalised linear models (GLMs, Nelder & Wedderburn, 1972). However, the sparseness of the data has the additional consequence that means of taxa are often near zero. Currently, marginal inference for multivariate abundance data is commonly implemented with Wald and score tests using generalised estimating equations (Section 2.2), as these can account for the discrete nature of the data, as well as covariance between species. However, Wald tests are known to lose power for counts when means are small, and score test can have poor power for unbalanced sampling designs, both common properties of multivariate abundance data. Gaussian copulas, in addition to modelling covariance, are a powerful method for marginal likelihood based inference. They allow for hypothesis testing of treatment effects, and model selection for important environmental predictors, free us from the limitations of GEE statistics. In Chapter 7 we propose Gaussian copulas as a method for marginal inference. We investigate how our method overcomes the limitations of existing inference methods and demonstrate superior power properties.

We conclude the introduction by taking a closer look at our motivating datasets.

1.1 Example datasets

We will demonstrate the methods proposed in this thesis with three diverse multivariate abundance datasets. Table 1.1 summaries some of the key properties of these data. The chosen datasets possess some of the common properties of multivariate abundance data. All of them have a high proportion of zeros (leading to strong discreteness), and a large number of variables relative to the sample size. The count datasets are overdispersed relative to the Poisson distribution (Figures 1.1b and 1.2).

The hunting spider data is included as it is a well known ecological dataset popularised in an important methodological paper (ter Braak, 1986). The bush regeneration data on the other hand has an unbalanced sampling design, which complicates analysis and impacts on the power of tests statistics for marginal hypothesis testing. The NZ native forest data are very sparse, high dimensional and ordinal, all of which present a challenge to analysis.

Throughout this thesis, as is commonly done in literature, we will use the term "species" to loosely mean the taxonomic level to which data are classified. But for some of these data (like the bush regeneration data) the variables are often not classified to species, but rather another taxonomic level.

Dataset	# of sites	# species	Proportion of zeros	Response
Bush regeneration	10	24	0.36	count
Hunting spider	28	12	0.46	count
NZ* native forest	964	1311	0.97	% cover(ordinal)

Table 1.1: Data summary: All datasets have high proportions of absences and a large number of species relative to sample size. *New Zealand.

1.1.1 Bush regeneration data

These data are part of a survey to assess the effect of vegetation restoration on invertebrate communities (Data were obtained from Anthony J. Pik at Macquarie University). Invertebrates were counted from pitfall traps at ten sites. Pitfall traps are holes dug in the ground, and all animals that fall in (and are trapped) are counted. Note that only two of the sites were control sites, with eight having undergone bush regeneration projects.

There were five pitfall traps per site, though one trap was lost. Invertebrates were classified to order (a level of phylogenetic classification of plants and animals, above family but below class) and aggregated across samples for analysis. A total of 24 orders were observed in the study. The primary question of interest is to test for a difference in the invertebrate communities between regenerated and control sites, as an indicator of success of the regeneration efforts.

There are no published ecological results for these data, though visual analysis (Figure 1.1) indicates a difference between regenerated and control sites for some species. These data are sometimes used for methodological work in ecological statistics; for example, Warton (2017) used this dataset to demonstrate that transformations cannot stabilise variance for small counts.

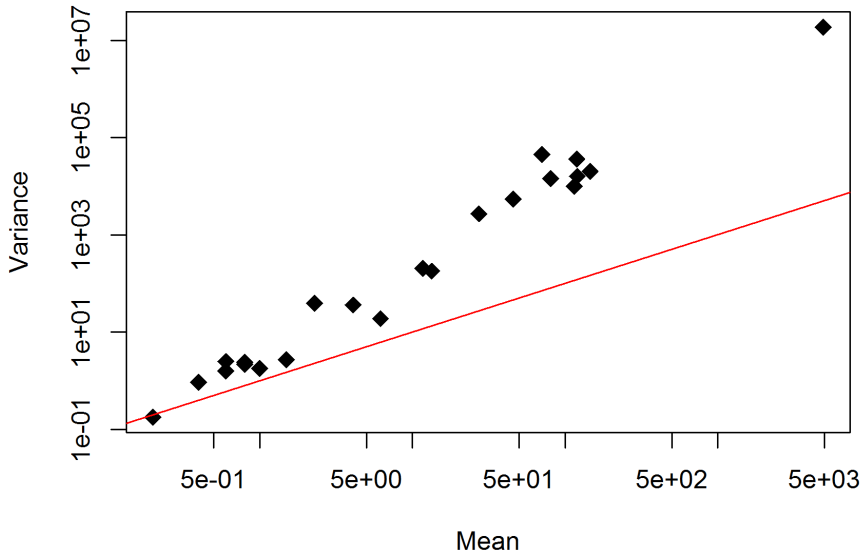
Treatment	Acarina	Araneae	Blattodea	Collembola	···	Seolifera
Control	21	12	3	1093	···	0
Reveg	70	1	0	580	···	0
Reveg	306	3	0	13541	···	0
Reveg	98	7	0	2809	···	0
Control	8	5	4	477	···	0
Reveg	112	13	1	7527	···	0
...	...	...	...	...	...	...
Reveg	320	10	0	5184	···	5

Table 1.2: Bush regeneration data (sub-sample): Counts of invertebrates at several sites. Some orders (*e.g. Blattodea*, *Seolifera*) are rare, while others (*e.g. Collembola*) are abundant. There are only two control sites in the data, and eight impacted sites.

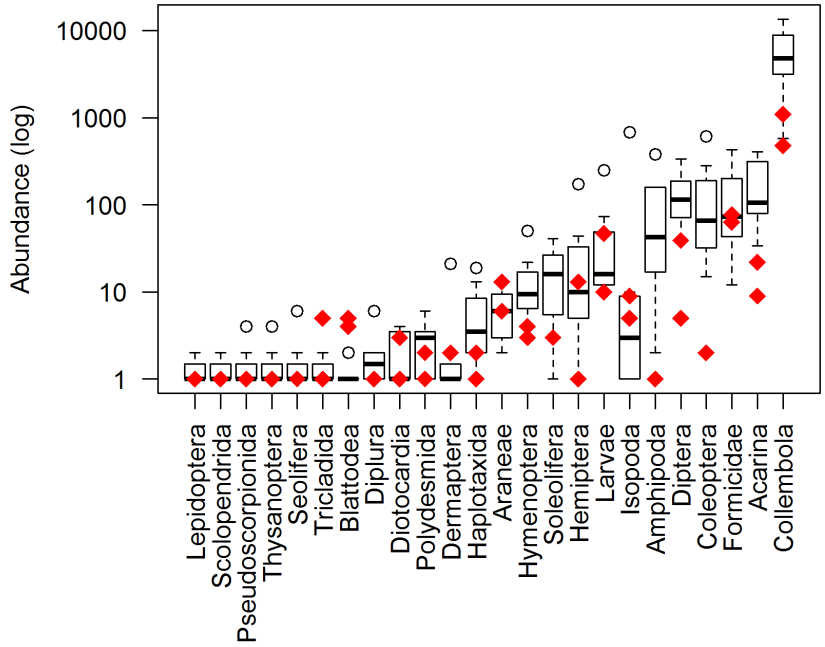
Table 1.2 contains a subset of the data. Invertebrates in this dataset are classified to order, rather than species. Ten classes are represented, with the most common being *Insecta* and *Arachnida*. We can see some orders (*e.g. Blattodea*) have very small counts with many absences, while others (*e.g. Collembola*) have very large counts. In Figure 1.1 (top) we have plotted the mean of each order against the variance. The counts appear overdispersed relative to the Poisson, with the variance exceeding the mean for most species. Figure 1.1 (bottom) shows the abundance of each order at regenerated sites (boxplot) overlayed with the abundance at the two control sites (red points). We can see that for some species (*e.g. Blattodea* and *Diptera*) there appear to be large differences in abundance between control and regenerated sites.

1.1.2 Hunting spider data

These data consist of counts of hunting spiders caught in pitfall traps, with 12 species found at 28 sites (van Der aart & Smeenk-Enserink, 1974). The primary aim of this study was to identify the main environmental factors associated with the distribution of the species studied. The data contain six covariates thought to be associated with spider abundance, namely: dry soil mass; percent cover of bare sand; percent cover of fallen leaves or twigs; percent cover of moss; percent cover of



(a)



(b)

Figure 1.1:

(a): Points are mean and variance of bush regeneration data by order, with red line at variance equals mean (Poisson assumption). The variance is larger than the mean for most species, implying overdispersion.

(b) Boxplot of (log of) abundance by species at regenerated sites, overlayed (in red) with (log of) abundance at the two control sites. Some orders, like *Blattodea* and *Diptera*, seem to be impacted by the regeneration while others are not.

herb layer and reflection of the soil surface with a cloudless sky.

Table 1.3 presents a subset of the data and two of the covariates (cover of bare sand, cover of fallen leaves or twigs). Species in the data are *Alopecosa accentuata (Alopacce), Alopecosa cuneata (Alopcune), Alopecosa fabrilis (Alopfabr), Arctosa lutetiana (Arctlute), Arctosa perita (Arctperi), Aulonia albimana (Auloalbi), Pardosa lugubris (Pardlugu), Pardosa monticola (Pardmont), Pardosa nigriceps (Pardnigr), Pardosa pullata (Pardpull), Trochosa terricola (Trocterr), Zora spinimana (Zoraspin).* Of the twelve species in this dataset, eleven are of the family *Lycosidae* (Wolf spider) with *Zora spinimana* being of the family *Miturgidae.* Of the wolf spiders, four species are of genus *Pardosa*, three are of the genus *Alopecosa*, two are of genus *Arctosa*, and one species of each of *Aulonia* and *Trochosa.* Species vary in terms of mean abundance, with some species (*e.g. Pardnigr*) having large means while other species (*e.g. Alopfabr*) are commonly absent. The plot of means against variances (Figure 1.2) shows overdispersion relative to the Poisson, with species having larger variance than their mean. Relationships between the species and two of the covariates (cover of bare sand and cover of fallen leaves or twigs, converted to binary variables) are shown in Figure 1.3. We observe relationships between some species abundances and both environmental predictors. For example, species *Alopfabr* and *Pardlugu* seem to respond to the presence of bare sand, while *Alopacce* and *Aluoalbi* each differ according to the presence of fallen leaves.

Previous analyses have found strong relationships between species abundance and environment (ter Braak, 1986), with patterns of association between species largely described by environmental variables (Peres-Neto et al., 2001; Hui et al., 2015a).

1.1.3 New Zealand native forest cover data

The dataset contains records of cover for 1831 forest plant taxa (most are classified to species level) at 1246 sites collected as part of the New Zealand Carbon Monitoring System, with the aim of advancing national-scale biodiversity reporting and monitoring. Data collection is lead by *Manaaki Whenua* landcare research in conjunction with Scion Research institute. These data included many types of plants, with the most common classification being Forb, Graminoid, Fern, Shrub and Tree.

Site	Bare sand	Fallen leaves	Alopacce	Alopfabr	Pardmont	···	Pardnigr
1	0	0	25	0	60	···	12
2	0	1.7918	0	0	1	···	15
3	0	0	15	2	29	···	18
4	0	0	2	0	7	···	29
5	0	0	1	0	2	···	135
6	2.3979	3.434	0	0	11	···	27
7	0	0	2	0	30	···	89
8	0	4.2627	0	0	2	···	2
···	···	···	···	···	···	···	···
28	3.434	0	15	14	6	···	0

Table 1.3: Hunting spider data (sub-sample): Counts of spiders at several sites. Some species (*e.g. Alopfabr*) are rare, while others (*e.g. Pardnigr*) are abundant.

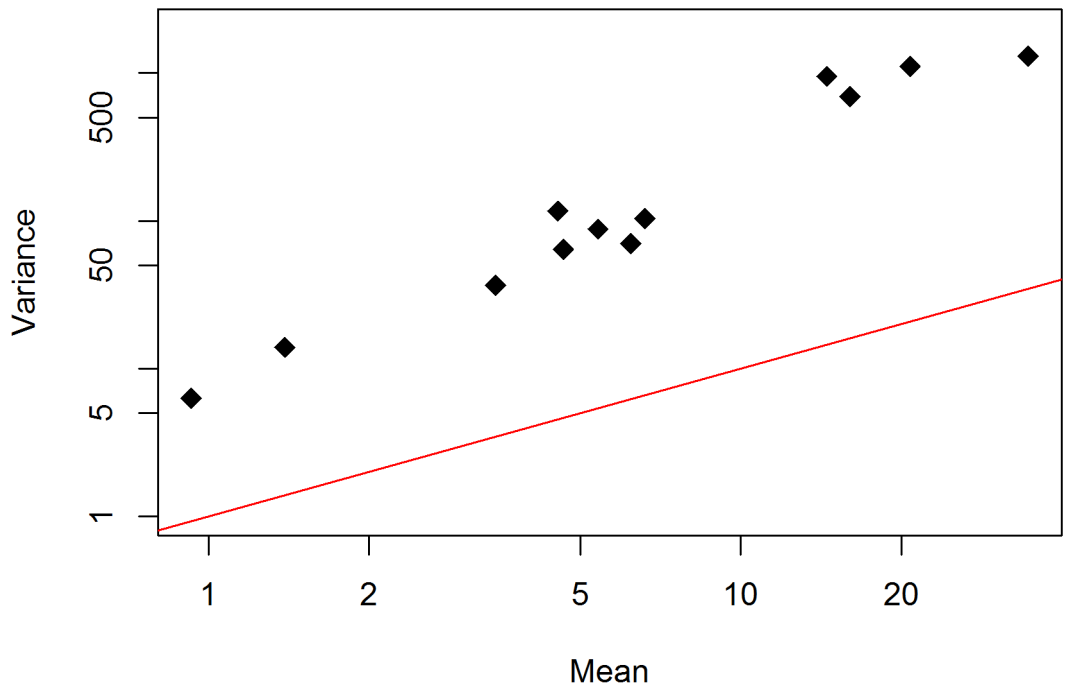


Figure 1.2: Points are mean and variance by species of hunting spider, with red line at variance equals mean (Poisson assumption). The variance is larger than the mean, implying overdispersion.

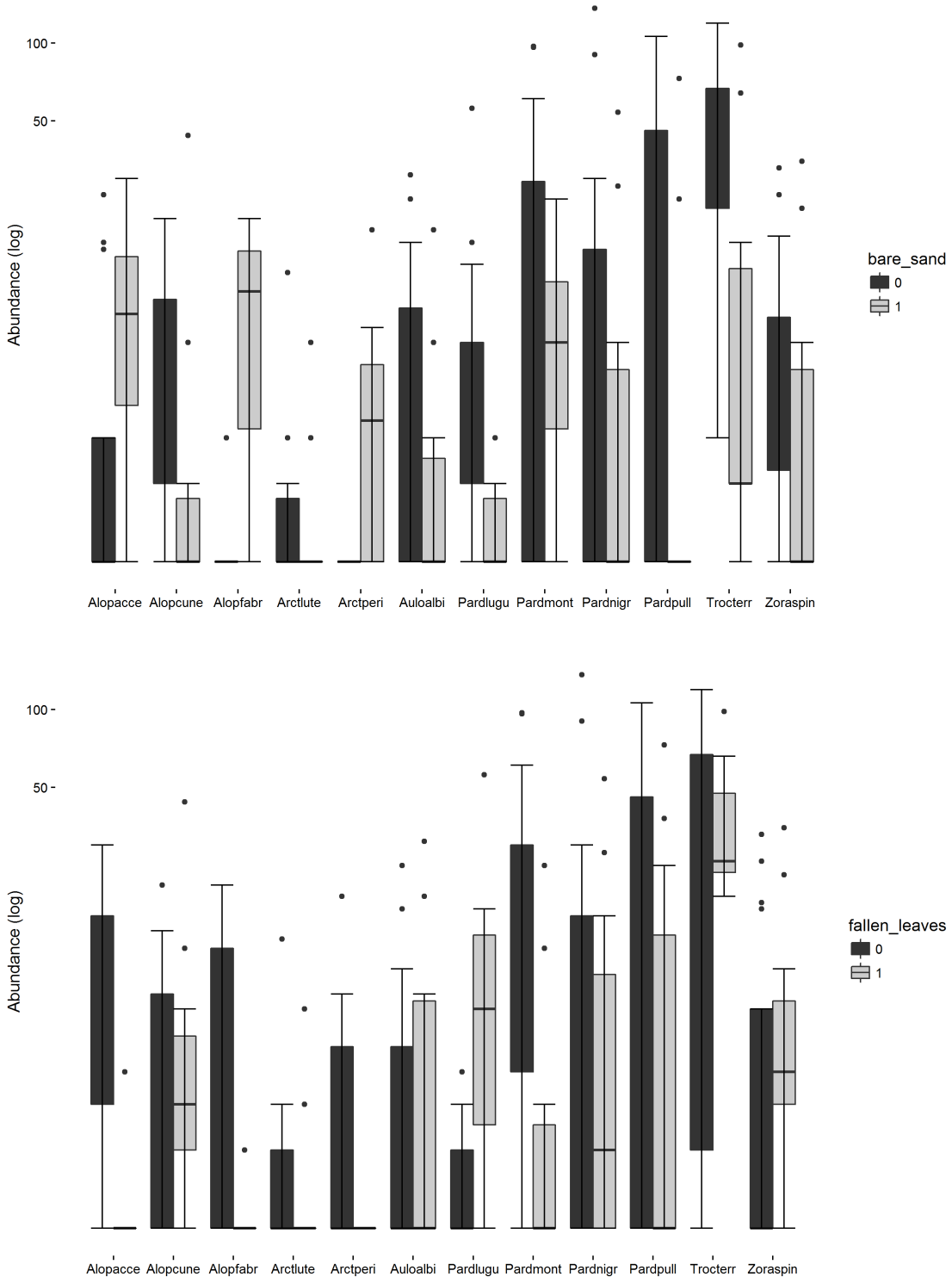


Figure 1.3:

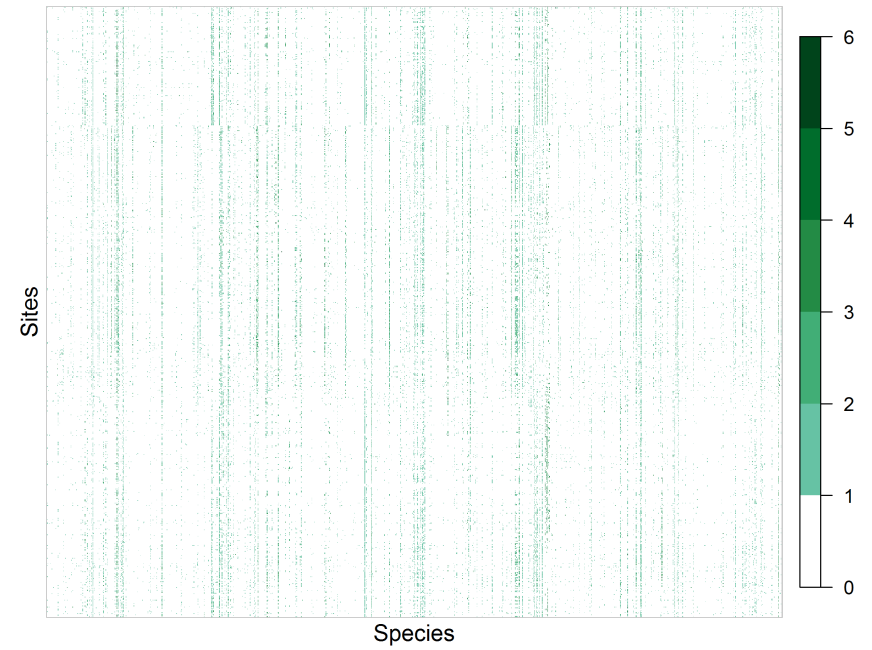
Top: Abundance by species of hunting spider and presence of bare sand. For several species (*e.g.* *Alopfabr* and *Pardlugu*) there seems to be an effect of bare sand.

Bottom: Abundance by species and presence of fallen leaves. Some species abundances (*e.g.* *Alopacce* and *Aluoalbi*) seem to be affected by the presence of fallen leaves.

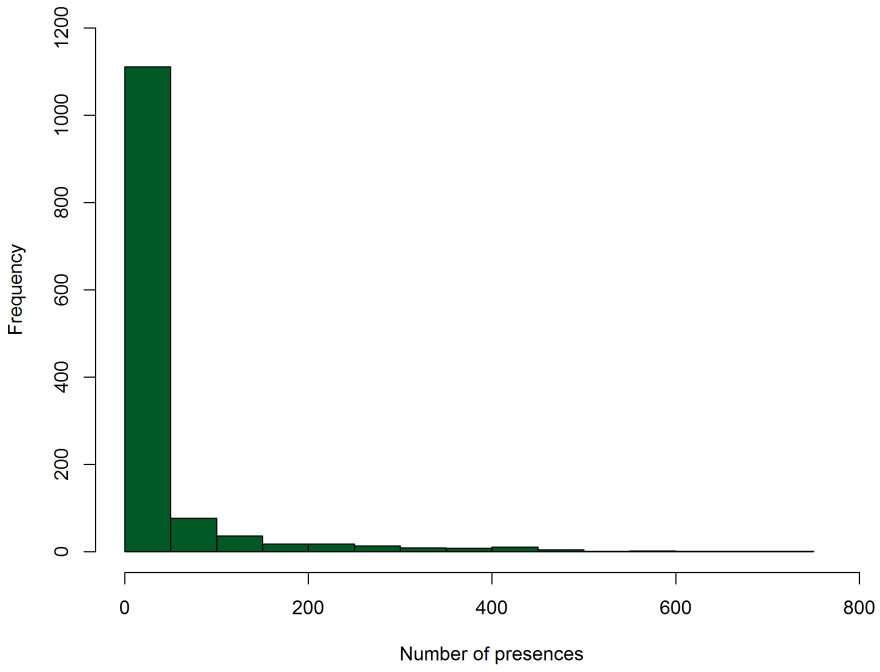
Analyses of these data have found that tree distributions are most strongly predicted by mean annual temperature and mean annual solar radiation (Leathwick, 1995). Additionally tree-ferns are found to interact with other species though their impact on nutrient cycling, organic matter accumulation and ground-level irradiance, often shading out tree seedlings (Brock et al., 2016). Analysis of traits found that leaf size and wood density are predictive of the forest phase: shaded understoreys, tree-fall gaps, treefern groves and clearings (Lusk & Laughlin, 2016).

A network of permanent 0.04-ha (20 $\times$ 20 m) plots are spread throughout New Zealand's indigenous forests and shrublands. Protocols for forest plot measurements are based on Allen (1993). Cover (in ordinal categories) was assessed for each species in several tiers at different heights. To obtain an estimate of total cover at a site, we took the maximum cover over all the tiers. We use a subset of these data which contains measurements at 964 sites identified as native forests, which reduced the number of species with at least one presence to 1311. The data additionally includes a number of covariates including altitude, aspect, slope, soil depth, drainage, ground cover, canopy height and cover, herbivore damage and fauna. In Figure 1.4a, we plot the raw data for all 1311 species, with white representing absences and darker colours representing higher cover categories (Table 1.4). We can see that most observations (97%) are absences, and only a few species are present in a large number of locations (Figure 1.4b). In Figure 1.5 we plot the Pearson correlations of species presences for the 123 species with 100 or more presences. Most species appear to be uncorrelated. For those that are correlated, the majority of strong correlations are positive.

These data have often been used to test and demonstrate models which find relationships between species and environment, or cluster species and sites. These analyses have found that the measured environmental variables explain the majority of the variation among the spider catches (ter Braak, 1986). The species can best be clustered into three clusters, or two latent variables. Sites 2224, 2628 were characterized by Arctosa perita (Arctperi) and to a lesser extent Alopecosa fabrilis (Alopfabr), while Pardosa lugubris (Pardlugu) strongly identified with sites 8, 1521 (Hui et al., 2015a).



(a)



(b)

Figure 1.4:

(a) Plot of raw NZ cover data. Absences are white, with darker colours for higher cover categories (Table 1.4). Most observations (97%) are absences.

(b) Number of presences by species at 964 sites. Most species are present in less than 50 sites.

Category	0	1	2	3	4	5	6
Cover	0	<1%	1-5%	5-25%	25-50%	50-75%	75-100%

Table 1.4: Cover categories for New Zealand native forest data. *i.e.* cover between 25% and 50% is recorded as category 4.

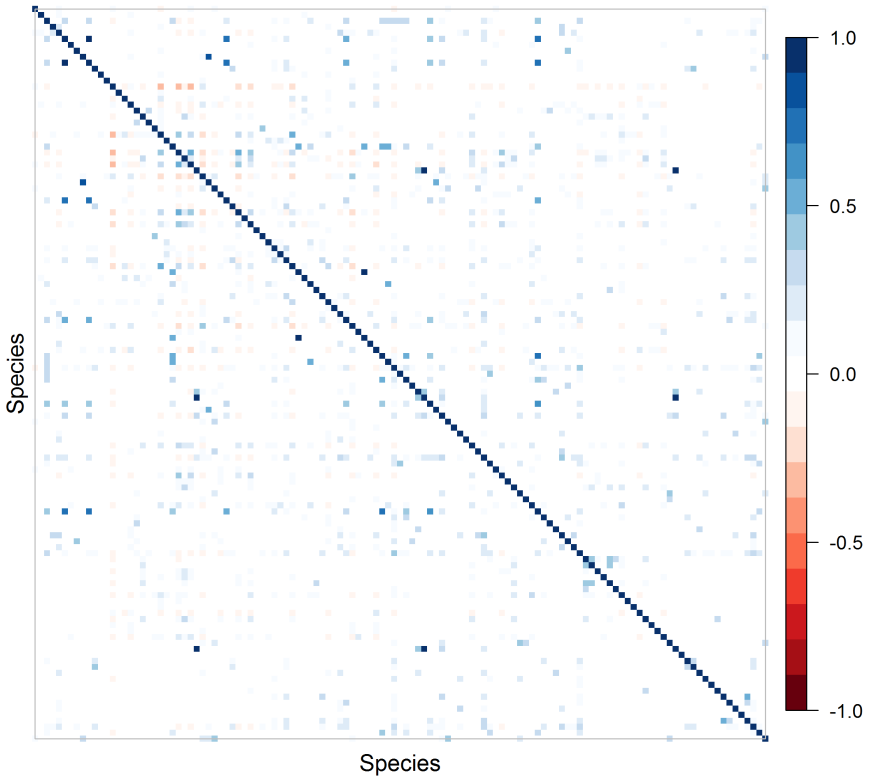


Figure 1.5: NZ cover data: Correlation of species presences for species with 100 or more presences. Most species are not highly correlated. Of those that are, most species correlations are positive.

Chapter 2

Model based analysis of multivariate abundance data

Historically, multivariate abundance data have been analysed with algorithmic methods (Legendre & Legendre, 2012). These methods first transform data, to try to account for the mean-variance relationship present, and then construct a pairwise dissimilarity matrix. Dissimilarities can be distance measures (such as euclidean distance) or, more generally, any measure of how dissimilar vectors are, like the Bray-Curtis dissimilarity (Bray & Curtis, 1957; Legendre & Legendre, 2012), which quantifies proportional dissimilarity. Bray-Curtis dissimilarity is not a distance as it does not obey the triangle inequality.

These dissimilarities are then modelled as a function of covariates. More recently, and in line with modern statistical practice, model based approaches have become more widely used to analyse these data. Model based approaches have many advantages to algorithmic methods, including better power properties, the ability to answer more diverse and interesting questions, and proven methods for checking assumptions (Ives & Helmus, 2011; Warton et al., 2015b). In this chapter we will review several model based methods most commonly used for analysing multivariate abundance data.

2.1 Models assuming independence

Due to the challenging nature of multivariate abundance data, it is historically common to assume a model where species are independent, and carry out inference based on this assumption. In ecology there is still a desire to carry out community level modelling (Ferrier & Guisan, 2006), and many interesting models with independence assumptions have been created for this purpose.

Many of these models are extensions of generalised linear models (GLMs; Nelder & Wedderburn, 1972), which are are commonly used for regression modelling of univariate non-normal data. A univariate response y is assumed to arise from an exponential family distribution F, with density function

$$f(y;\mu,\phi) = \exp\left\{\frac{y\mu - b(\mu)}{a(\phi)} - c(y,\phi)\right\}.$$

The canonical parameter μ, which for count and binomial distributions is the mean, is related to covariates $X = (x_1, ..., x_K)$ via a known link function

$$g(\mu_{ij}) = \beta_{0j} + X_i^T\beta_j,$$

for site $i = 1, \cdots N$ and species $j = 1, \cdots P$. This specification includes as special cases many well known discrete distributions, including the Bernoulli distribution for modelling binary data and the binomial, Poisson and negative binomial distributions for counts. There are several parameterisations for the negative binomial distribution, in this thesis we use the definition found in McCullagh & Nelder (1989), such that the variance $V(\mu) = \mu + \phi\mu^2$.

Most simply, one can model species separately as a function of environmental covariates using GLMs (*e.g.* Yee & Wild, 1996; Austin, 2002). These models assume individual species are distributed according to a distribution F_j, with species-specific dispersion parameter ψ_j, where the mean μ_{ij} depends on covariates X though species specific coefficients β_j, and a link function $g(\cdot)$. Letting the vector θ contain all the

model parameters, we have response Y_{ij} at site i for taxonomic group j, where

$$\begin{aligned} Y_{ij} &\sim F_j(\mu_{ij}, \psi_j) \\ g(\mu_{ij}) &= \beta_{0j} + X_i^T \beta_j \\ L(\theta) &= \prod_{j=1}^{P} \prod_{i=1}^{N} f_j(\mu_{ij}, \psi_j). \end{aligned}$$

It is important to highlight that these models are designed to model and predict for one species, and multivariate inference based on these models does not take into account any correlation between species, other than through common responses to known covariates.

An interesting extension of GLMs are finite mixture of regression models. These allow for community level modelling of multi-species data, by assuming all species can be classified into a small number groups (species archetypes), according to their environmental response (Dunstan et al., 2011; Hui et al., 2013). The model for $G \ll P$ archetypes is

$$\begin{aligned} Y_{ij}|g &\sim F_j(\mu_{ijg}, \psi_j) \\ g_j(\mu_{ijg}) &= \beta_{0j} + X_i^T \beta_g \\ L(\theta) &= \sum_{g=1}^{G} \pi_g \prod_{j=1}^{P} \prod_{i=1}^{N} f_g(\mu_{ij}, \psi_j). \end{aligned}$$

Here f_g is the model for the gth archetype, $g = 1, \ldots, G$, and π_g, (with $\sum_{g=1}^{G} \pi_g = 1$), are the mixing proportion of species whose mean response is governed by archetype g. This modelling framework allows species in the same archetype to have a shared response to covariates β_g, rather than individual species responses β_j. This formulation borrows strength across species of the same archetype to improve predictive performance, particularly for rare species.

Another mixture model approach (Foster et al., 2013) aims to classify sites in environmental space, by the pattern of species that occur (or are absent) there. One can also use mixture models to simultaneously group by species and site (Pledger & Arnold, 2014), with or without environmental covariates. In all these models, species are assumed to be independent conditional on covariates and group membership.

A different approach is to assume a joint structure for the response of individual species to environmental covariates. The method of Ovaskainen & Soininen (2011) assumes species-specific coefficients β_{jk}'s are distributed according to a multivariate Gaussian distribution for different species, while the VGAM method (Yee, 2004; Yee et al., 2010) models coefficients by reducing the rank of the matrix of β_{jk}'s.

One can also include trait information, as well as environmental covariates, and interactions of the two, as predictors in models (Pollock et al., 2012; Jamil & ter Braak, 2013; Brown et al., 2014; Warton et al., 2015c). These models aim to explain how environmental covariates act on species abundance, through their interaction with traits.

All these methods go some way to explaining correlation between species as a function of environment, site, traits, and membership of a cluster, but species are assumed to be independent conditional on these explanatory variables. These models fail to account for many sources of correlation between species, including unobserved covariates and species interactions. This leads to a loss of power at detecting truly important covariate effects when it comes to marginal inference (Section 7.1.1). In addition, these models, by definition, cannot be used to investigate patterns of correlation between species, as the correlation is not modelled.

2.2 Generalised estimating equations

2.2.1 Description

Generalised estimating equations (GEEs; Liang & Zeger, 1986; Zeger & Liang, 1986) are commonly used for marginal inference for correlated non-normal response variables. Over the past decade, GEEs have emerged as a commonly applied too for studying multivariate abundance data (Warton et al., 2015b), as they are straightforward to fit with existing software, and enable marginal inference, which can answer many of the questions of interest to ecologists. GEEs don't explicitly model the covariance structure of the data, but rather treat it as a nuisance parameter which is accounted for during inference on the marginal effects.

The starting point for GEEs are GLMs (defined in Section 2.1). For GLMs, model parameters are estimated by solving the score equations (estimating equations)

$$0 = \sum_{i=1}^{N} D_i v(\hat{\mu}_i)(y_i - \hat{\mu}_i), \tag{2.1}$$

where D_j is a vector whose kth element is $\partial\hat{\mu}_i/\partial\beta_k$, and $v(\hat{\mu}_i)$ is the variance as a function of the mean. GEEs extend the concept of estimating equations to correlated data, without the need to specify a joint model for the data. They introduce a correlation matrix R, such that the marginal covariance of the responses is parameterised as $V_i = A_i^{1/2} R A_i^{1/2}$, where A_i is a diagonal matrix of variances $v(\hat{\mu}_i)$. The estimating equations thus become

$$0 = \sum_{i=1}^{N} D_i V_i (y_i - \hat{\mu}_i). \tag{2.2}$$

When the correlation matrix is the identity ($R = I$), solving Equation (2.2) is equivalent to solving Equation (2.1), these are called independence estimating equations (Hilbe et al., 2003). The inclusion of a more general R can take correlation between variables into account for estimation or inference. For example, for panel data, R might have a compound symmetry structure, or autoregressive for temporally correlated data. Iteratively reweighted least squares is used for estimation if a closed form solution does not exist (McCullagh & Nelder, 1989).

2.2.2 Inference

To carry out inference, Wald and score tests can be used, as they only require an estimate of the covariance matrix of parameters. A naive estimate assumes that the correlation structure is correctly specified, and is given by

$$\hat{var}(\hat{\beta}) = \gamma_R^{-1} = \left(\sum_{i=1}^{N} D_i V_i^{-1} D^T \right)^{-1}. \tag{2.3}$$

A more robust estimate of the covariance matrix is given by the sandwich estimator (Liang & Zeger, 1986; Zeger & Liang, 1986)

$$\hat{var}_s(\hat{\beta}) = \gamma_R^{-1} \gamma \gamma_R^{-1}, \tag{2.4}$$

where

$$\gamma = \sum_{i=1}^{N} D_i V_i^{-1} (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)^T V_i^{-1} D^T.$$

An alternate, more efficient robust estimator of covariance (Pan, 2001b) can be constructed using a moment based estimator of $\hat{R}$ with Pearson residuals, so letting

$$\gamma = \sum_{i=1}^{N} D_i V_i^{-1} A_i^{1/2} \hat{R} A_i^{1/2} V_i^{-1/2} D^T \tag{2.5}$$

in Equation (2.4). This estimator is consistent when the mean-variance relationship is correctly specified and there exists a common correlation matrix for all observations.

In the context of multivariate abundance data, the correlation matrix specifies dependence between species. *A priori*, we might assume all pairwise species correlations are potentially different, and estimate an unstructured matrix. However, when sample size is small, this matrix cannot be estimated reliably because of the large number of correlation parameters relative to the number of observations, and some simplification of the structure is required. For GEEs, a regularised sandwich estimator (Warton, 2011) has been proposed in this context, as a way to obtain an estimable covariance matrix, with which to carry out inference. This estimator shrinks the estimate of the covariance matrix towards the identity matrix, to ensure it is positive definite.

Due to lack of an explicit likelihood function for GEEs, likelihood ratio statistics and information criteria can not be used for marginal inference (Rotnitzky & Jewell, 1990), unless independence between variables is assumed. Likelihood free marginal inference for parameters can be carried out using the GEE approach, as covariance between variables can be estimated, however inference not based on likelihoods can have less desirable properties (Section 7.1.2). Model selection for GEEs can also be carried out using quasi-information criteria, which make use of covariances (Pan, 2001a; Cantoni et al., 2005; Wang & Qu, 2009; Wang et al., 2012a; Cho & Qu, 2013). The quasi likelihoods used in these criteria are analogous to assuming independence between species, and hence these model selection criteria suffer from the same properties as inference based on assuming independence (Section 7.1.1).

As variance parameters are treated as a nuisance, it is not natural to model the patterns in covariance between variables with GEEs (Chapter 4).

2.3 Hierarchical models

Methods which explicitly model covariance between species have recently been developed. These are generally hierarchical models (*e.g.* Gelman & Hill, 2006). There are several incarnations of these models, but commonly the data are assumed to arise from a particular distribution, often from the exponential family, with some mean parameter μ_{ij}, and species-specific dispersion parameters ψ_j. So, as with GLMs,

$$Y_{ij} \sim F(\mu_{ij}, \psi_j).$$

The mean in then modelled as

$$g(\mu_{ij}) = X_i^T \beta_j + \epsilon_{ij},$$

where $g(\cdot)$ is a link function, X_i is a vector of covariates for the ith observation and β_j is a vector of species specific coefficients. The error ϵ_i is assumed to be multivariate Gaussian with some correlation matrix $\epsilon_i \sim N(0, \Sigma)$. The covariance matrix Σ here describes the residual covariance between species on the latent scale.

As we have little information about how species are correlated, it is natural that the covariance matrix is unstructured, as for example in Pollock et al. (2014). In order to estimate an unstructured covariance matrix, which has $(P-1)P/2$ parameters, a lot of data are required. As discussed in Section 1.1, this is not generally available for multivariate abundance data. Recently, some hierarchical models have overcome this limitation, by imposing some structure on Σ, to reduce the number of variables requiring estimation. One option is to assume that all pairs of species have the same correlation (Jamil & ter Braak, 2013), however this is highly restrictive and ecologically implausible. Reducing the dimension of the covariance matrix without overly restrictive assumptions can be done using latent variables (Section 4.1.1: Walker & Jackson, 2011; Hui et al., 2015a; Ovaskainen et al., 2016; Warton et al., 2015a; Hui, 2016). These models are difficult to fit, and are generally estimated using Markov chain Monte Carlo, or more recently variational approximations (Hui et al., 2016).

2.4 Conditional and marginal models

We have discussed several approaches for modelling multivariate abundance data, some of which are marginal (GEEs) while others are conditional (hierarchical). It is important to make a distinction between such models, as they have different interpretations and characteristics.

Latent variables (and random effects) in conditional models govern the covariance between variables as well as overdispersion characteristics, and can produce unintended artefacts. For example, assuming a bivariate Poisson distribution such that $\log E(y_{ij}|u_i) = \mu_j + \gamma u_i$, γ determines both the covariance between variables as well as overdispersion of each margin (Murray et al., 2013). For count data this can be overcome by using a negative binomial distribution marginally, to model overdispersion, leaving the latent variables to model only covariance (Ovaskainen et al., 2016; Hui, 2016; Hui et al., 2016).

Interpretation of parameters in conditional models is by nature conditional on random effects or latent variables, and distinct from interpretation in marginal models. For example, in a log linear model, the marginal mean is given by $E(Y) = \exp(\mu_j + \sigma^2/2)$, while in a similarly defined marginal model the mean is $E(Y) = \exp(\mu_j)$, giving μ_j a quite different interpretation marginally.

In this thesis we will focus on marginal modelling. Marginal modelling of multivariate abundance data is most commonly carried out with GEEs. We will introduce copulas as an alternative marginal model for multivariate abundance data (Chapter 3). Unlike GEEs, copula models specify a likelihood, making likelihood based inference available. This can greatly improve the power of inference, including hypothesis testing and model selection (Chapter 7). In addition, covariance between species are explicitly modelled by copulas. This allows us to gain insight into patterns of correlation by using covariance models (Chapter 4).

Chapter 3

Copulas for discrete data

In order to model multivariate abundance data, we have to solve two sets of problems simultaneously. Firstly, we need a way to model multivariate discrete data in a way which allows flexible (positive and negative) correlations between variables. Secondly, due to the small sample sizes and sparse data, we also do not have enough data to estimate the full set of correlations between species, and so must introduce some structure into the covariances.

This chapter will discuss copula models, with a focus on Gaussian copulas for discrete data. These will allow us to flexibly model discrete multivariate abundance data. Later chapters will focus on modelling covariances parsimoniously.

3.1 Copulas

Copulas are a flexible class of models which allow the modelling of data from any set of marginal distributions, with the covariance structure of any multivariate distribution. Copula modelling is rooted in Sklar's Theorem (Sklar, 1959). This states that the joint cumulative distribution function of a P-variate random variable, $H(y_1, y_2, \cdots, y_P)$, $j = 1, 2, \cdots, P$, can be written in the form

$$H(y_1, y_2, \cdots, y_P) = C(F_1(y_i), F_2(y_2), \cdots, F_p(y_P)), \tag{3.1}$$

where C and F_j are uniquely determined when H is known and continuous. Here F_j is the jth marginal distribution and $C : [0,1]^P \to [0,1]$ is known as a copula. In

other words, any multivariate distribution is a copula model, provided we have the correct specification for the marginal distributions and copula. From a modelling perspective, the main appeal of copulas is that the covariance structure can be specified independently of the marginal distributions, making them very flexible.

Copula models can be built assuming parametric forms for the $F_j \in F_\theta$ and $C \in C_\rho$, or semi-parametrically or non-parametrically, where either the F_j or C or both do not assume a parametric form (*e.g.* Genest et al., 1995; Shih & Louis, 1995; Deheuvels, 1979). They have a long history in econometrics (Cherubini et al., 2004) as well as engineering (Genest & Favre, 2007; Favre et al., 2004; Salvadori & De Michele, 2004) but are only sporadically used in other fields.

We start by discussing parametric copulas with continuous margins, and how they can easily be derived from multivariate distributions. We then introduce copulas with discrete margins, which can be understood as a latent variable model. We go on to define the Gaussian copula with discrete margins, for which estimation methods are discussed. We compare these models to generalised estimating equations, which are commonly used to model multivariate abundance data, and discuss how we will extend existing copula models to deal with multivariate abundances.

3.2 Parametric copula models

For a continuous random variable, the joint density $h(\cdot)$ can be derived by differentiating Equation (3.1) to obtain

$$h(y_1, y_2, \cdots, y_P) = c(F_1(y_i), F_2(y_2), \cdots, F_P(y_P)) \prod_{j=1}^{P} f_j(y_j),$$

where F_j and f_j are the marginal distributions and densities, and $c(u) = \partial C(u)/\partial u$ is the copula density. Parametric copulas specify a parametric distribution for each of the $F_i(\cdot)$ as well as the copula density $c(\cdot)$. One way to construct valid copulas is from existing multivariate distributions using the probability integral transform theorem (Nelsen, 1999).

3.2.1 Gaussian copulas

A valid copula density must have $C : [0,1]^P \to [0,1]$ with uniform margins. To derive a Gaussian copula, we start with a P-variate Gaussian distribution with zero mean, unit variance, and correlation matrix R,

$$\mathbf{\Phi}_P(z_1, z_2, ..., z_P; R) = \frac{1}{(2\pi)^{k/2}|R|^{1/2}} \exp\left(-\frac{1}{2} z^T R^{-1} z\right).$$

Then using the probability integral transform we have, for a univariate Gaussian, $\Phi(z_i) = u_i$, where $u_i \in [0,1]$ and $u_i \sim \text{Unif}(0,1)$. So we construct a Gaussian copula

$$C_G(u_1, u_2, ..., u_p) = \mathbf{\Phi}_P(\Phi^{-1}(u_1), \Phi^{-1}(u_2), ..., \Phi^{-1}(u_P), R)$$

The corresponding copula density is then given by

$$c_G(u_1, u_2, ..., u_p) = |R|^{-1/2} \exp\left(-\frac{1}{2} z^T (R^{-1} - I) z\right), \tag{3.2}$$

where $z_j = \Phi^{-1}(u_j)$.

Data from any continuous distribution can be transformed, *via* a probability integral transform, such that it is marginally standard Gaussian. The key assumption in a Gaussian copula model is that these marginally Gaussian variables are jointly multivariate Gaussian.

3.3 Copulas for discrete data

In the case of discrete data, the copula distribution in Equation (3.1) is not unique, however it is uniquely defined on the Cartesian product of the ranges of the marginal distribution functions (Genest & Neslehova, 2007). This non uniqueness does not prevent the use of parametric copulas for modelling discrete data. In the case where Y_j are all discrete, the copula density is found by obtaining the 2^P finite differences

$$P(Y = y) = \sum_{i_1=0,1} \cdots \sum_{i_P=0,1} (-1)^{i_1+\cdots+i_m} C(F_1(y_1 - i_1), \cdots, F_P(y_P - i_P)). \tag{3.3}$$

There are 2^P term in this equation, and so to compute the likelihood for a sample of size N, the copula distribution must be evaluated at $N \times 2^P$ points. Alternately, we can write the copula as a latent variable model, where latent variables u have a

copula distribution $c(u)$, and for each y_j, $f(y_j|u_j) = I(F_j(y_j^-) \leq u_j < F_j(y_j^-))$. The joint distribution of y and u is given by

$$f(y,u) = f(y|u)f(u) = \prod_{j=1}^{P} I(F_j(y_j^-) \leq u_j < F_j(y_j))c(u).$$

The required marginal distribution of y is then given by

$$\begin{aligned} f(y) &= \int f(y,u)du = \int \prod_{j=1}^{P} I(F_j(y_j^-) \leq u_j < F_j(y_j))c(u)du \\ &= \int_A c(u)du, \end{aligned}$$

where $A_i = \cap_j \left[(F_{ij}(y_{ij}^-|\beta_j,\psi_j), F_{ij}(y_{ij}|\beta_j,\psi_j)\right]$. This is equivalent to Equation (3.3) (Smith & Khaled, 2012). There are a number of examples of parametric copulas used for discrete data, see Nikoloulopoulos (2013a) for full discussion. We are aiming to model multivariate abundance data, where pairs of species can potentially have either positive or negative dependence, and correlations for each pair are potentially different. Only the elliptical (including the Gaussian copula) and Vine copulas are able to model a wide range of dependence, including positive and negative dependence, in a way that allows the pairwise dependence between variables to be different for each pair of variables.

3.4 Gaussian copula with discrete margins

For discrete Gaussian copulas we employ Equation (3.4) with the Gaussian copula density in Equation (3.2), to obtain the likelihood

$$L(y|\beta,\psi,\theta) = \prod_{i=1}^{N} \int_{A_i} |R_\theta|^{-1/2} \exp\left(-\frac{1}{2} z_i^T (R_\theta^{-1} - I) z_i\right) du_i, \tag{3.4}$$

where $z_{ij} = \Phi^{-1}(u_{ij})$ and $A_i = \cap_j \left[(F_{ij}(y_{ij}^-|\beta_j,\psi_j), F_{ij}(y_{ij}|\beta_j,\psi_j)\right]$, with $i = 1,\ldots,N$ observations. A_i is a hypercube of all values on the copula scale which, when discretised, give us the observed Y.

To maximise the copula likelihood, it is necessary to first approximate the required rectangular integrals (3.4), and then maximise the resulting approximate likelihood. The latter is generally done using standard numerical optimisation methods like quasi-Newton algorithms (Nikoloulopoulos, 2013a).

To approximate the integrals for low dimensional problems, or in cases where the likelihood can be factored in a small dimension, such as with clustered data with a small number of observations per cluster, it is possible to use deterministic approximations to the required rectangle integral (Joe, 1995; Miwa et al., 2003; Craig, 2008), which are then maximised numerically.

For larger dimensional problems, where deterministic approximations are computationally infeasible, it is common to use sampling methods to approximate the integrals, followed by numerical optimisation. This procedure is effective when the likelihood cannot be factorised to a smaller dimension, but correlations can be parametrised by a small number of parameters, as in time series and spatial problems. Examples of sampling methods include the randomised quasi-Monte Carlo (Genz & Bretz, 2002), as well as importance sampling methods (Masarotto & Varin, 2012).

In very high dimensional problems, when maximising the full likelihood is not feasible, a two stage method can be used. Here parameters belonging to marginal distributions are estimated assuming independence, and then correlations are estimated conditional on the marginal parameters. Letting $l_j(\theta_j)$ be the jth marginal log likelihood, and $l(\theta, R)$ the complete log likelihood, the two stage method solves

1. $\hat{\theta}_j = \arg\max_{\theta_j} l_j(\theta_j)$ for $j = 1, \cdots, P$, and
2. $\hat{R}_j = \arg\max_R l(\hat{\theta}, R)$.

A further method of approximating the likelihood is the method of inference functions for margins (Xu, 1996), which employs lower dimensional margins. This method is used in high dimensional problems, when estimation of all covariance parameters simultaneously is infeasible, even when marginal parameters have been estimated first. For the Gaussian copula, for example, each covariance matrix parameter $\rho_{j,k}$ appears only in the joint marginal distributions of variable j and k. Writing the bivariate j, k margin as $l_{j,k}(\theta_j, \theta_k, \rho_{jk})$, this method would proceed by solving for all pairs j and k

$$(\hat{\theta}_j, \hat{\theta}_k, \hat{\sigma}_{j,k}) = \underset{\theta_j,\theta_k,\rho_{j,k}}{\arg\max}\, l_{j,k}(\theta_j, \theta_k, \rho_{jk}).$$

This maximisation can be carried out jointly, or sequentially as in the two stage method. The estimators (θ, R) are consistent and in many cases efficient (Xu, 1996; Joe, 2005)

3.5 Modelling multivariate abundance data with copulas

Gaussian copulas with discrete margins provide a flexible framework for modelling multivariate discrete data. They have many desirable properties in this context. Any set of marginal distributions can easily be combined to form a copula model, with the only requirement being a well specified density and cumulative distribution functions for each margin. We could therefore easily model some species with presence/absence data, with ordinal or count data for others. Additionally, Gaussian copulas allow very flexible correlation structures to be modelled. In later chapters we will see that Gaussian copulas also allow us to easily specify parsimonious but flexible correlations. In this section we will briefly compare copulas with generalised estimating equations, which are commonly used for marginal inference for multivariate abundance data (Section 2.2). We will then describe an existing framework for multivariate generalised linear models, which we will in later chapters extend to model multivariate abundance data.

3.5.1 Comparison to generalised estimating equations

There is a close relationship between copulas and marginal models estimated with GEEs. Both modelling frameworks assume the same marginal distributions,

$$Y_{ij} \sim F_j(\mu_{ij}, \psi_j)$$
$$g(\mu_{ij}) = \beta_{0j} + X_i^T \beta_j,$$

and the same model under an independence assumption, with likelihood specified by

$$L(\theta) = \prod_{j=1}^{P} \prod_{i=1}^{N} f_j(\mu_{ij}, \psi_j). \tag{3.5}$$

In addition, maximum likelihood estimates using independence estimating equations (Hilbe et al., 2003) are equivalent to parameter estimates under a two stage estimating procedure (Section 3.4) for copulas, as both procedures estimate marginal models separately for each species. In practice this is done by fitting GLMs marginally to each species, and this procedure is commonly used when implementing GEEs for multivariate abundance data (Wang et al., 2012b).

GEEs for correlated data do not explicitly specify a likelihood, while copulas do, for example the Gaussian copula likelihood is given in Equation (3.4). This is the main advantage of copulas over GEEs, as it allows likelihood based inference like likelihood ratio tests and information criteria. In addition, copula models can accommodate mixed data types, while GEEs generally do not. Copula models explicitly model correlation, and so can be used to investigate patterns of covariance between species (Chapter 5), while GEEs treat correlations as nuisance.

3.5.2 Multivariate GLMs using Gaussian copulas

Many of the models in Chapter 2 are extensions of generalised linear models to multivariate outcomes. It is common to use GLMs as a basis for modelling multivariate abundance data marginally, as it accounts for the mean-variance relationship commonly observed in discrete data.

GLMs have previously been extended to multivariate data with Gaussian copulas (Song et al., 2009). Masarotto & Varin (2012) establishes a framework for modelling correlated clustered, longitudinal, temporal and spatial data using the likelihood specification in Equation (3.4), where the marginal distributions F_j are assumed to arise from a distribution in the exponential family, as in traditional GLMs.

The models described in Song et al. (2009) are highly multivariate, in the sense that there are many correlated variables being modelled, however the covariance between

response variables is parametrised with a small number of variables, which does not grow with the number of response variables, such as spatial or temporal covariance functions for example. This type of model can be estimated by obtaining an estimate of the full likelihood with sampling approximations, and then numerically maximising this estimate. This is possible as there are relatively few parameters to maximise over.

For multivariate abundance data, the number of variables in the correlation matrix is quadratic in the number of response variables. Maximising the required likelihood is difficult, but can be done (for moderate P) using approximate methods, if there is enough data to estimate all the parameters. For multivariate abundances, we often have few observations relative to the number of response variables, and so an unstructured covariance matrix is not estimable. In Chapter 4 we look at how covariance modelling techniques can be used to reduce the number of parameters in these models while making few and ecologically defensible assumptions. We will then, in Chapter 5, describe a novel algorithm for the estimation of these models, and demonstrate how they can be applied to model multivariate abundance data.

Chapter 4

Covariance modelling of multivariate data

As well as being highly multivariate, abundance data often does not have many observations relative to the number of variables. So while we would like to model an unstructured covariance matrix, where pairwise correlations are free to vary independently, this specification is generally not estimable from the data. In order to define a parsimonious model, some structure must be added to the covariance matrix, we will refer to this as covariance modelling (Pourahmadi, 2013).

Covariance modelling includes design driven covariance structures like compound symmetry or autoregressive covariances, as well as data driven covariance models like latent variable and graphical models. In this chapter we will focus on data driven covariance models, as these can induce parsimony in multivariate models in a flexible way and with few assumptions. Additionally, they are a method to uncover interesting patterns in dependence of multivariate data.

4.1 Covariance modelling of Gaussian data

Data driven models for covariance give us valuable information about the structure of multivariate data, when there are a large number of response variables, and the literature on such tools for Gaussian data is quite advanced. Two methods of

particular interest are latent variables models and graphical models.

4.1.1 Latent variable models

Latent variable models assume covariance between response variables is driven by a shared response to several unobserved latent variables. In the context of ecology, these are often interpreted as unobserved environmental covariates (Warton et al., 2015a). The simplest form of latent variable model is a factor analysis (Everitt, 1984). Here the response Y (of dimension P) is independent Gaussian with diagonal covariance matrix Ψ, conditional on a latent Gaussian variable $Z \sim N(0, I)$ with dimension $Q \ll P$. We can write

$$Y = \mu + \Lambda Z + W,$$

where $W \sim N(0, \Psi)$ independently of X, and Λ is a matrix of factor loadings. If we know Z, then this is a linear regression, but Z is unobserved. To find the marginal mean and covariance of Y we observe

$$\begin{aligned} E(Y) &= E(\mu + \Lambda Z + W) \\ &= \mu + \Lambda E(Z) + E(W) \\ &= \mu, \\ Var(Y) &= E[(\mu + \Lambda Z + W - \mu)(\mu + \Lambda Z + W - \mu)^T] \\ &= E[(\Lambda Z + W)(\Lambda Z + W)^T] \\ &= \Lambda E(ZZ^T)\Lambda^T + E(WW^T) \\ &= \Lambda\Lambda^T + \Psi. \end{aligned}$$

Covariance between variables is therefore a sum of shared responses to the latent variables $\Lambda\Lambda^T$, where each variable has an additional variance component $\Psi_{j,j}$. Factor analysis induces structure into the covariance matrix, which can give us important information about how variables are correlated. We can plot factor loadings $\Lambda_{j,q}$, with $q = 1, \cdots, Q$, to look for patterns among variables, where variables close to one another are highly correlated, and respond similarly to the latent variables. We can also plot scores $Z_{i,q}$, to look for which sites cluster together, and are therefore correlated.

Another advantage of latent variable methods is that they can significantly reduce the number of variables for estimation in the covariance matrix, with minimal assumptions about the covariance between variables. The number of variables in a factor analysis model, after ensuring identifiability (Anderson, 1962) is the number of variables in Λ (PQ), minus identifiability constraints $Q(Q-1)/2$, plus P, the number of elements in the diagonal matrix Ψ. So $K_{FA} = P(Q+1) - Q(Q-1)/2$, while the number of elements in an unstructured covariance matrix is $K_{UN} = P(P-1)/2$. Table 4.1 shows that even for moderate P, a factor analysis with up to six latent factors will substantially reduce the number of parameters that need to be estimated. This is due to the number of variables in a factor analysis being linear in P, rather than quadratic. This is particularly important for small sample sizes, as is common in multivariate abundance data.

Number of variables (P)	20	20	20	40	40	40	60	60	60
Number of factors (Q)	2	4	6	2	4	6	2	4	6
Unstructured	190	190	190	780	780	780	1770	1770	1770
Factor Analysis	59	94	125	119	194	265	179	294	405

Table 4.1: Number of parameters in unstructured and latent variable covariance matrices. The number of parameters in a factor analysis can be orders of magnitude lower than an unstructured covariance matrix.

Factor analysis models are commonly estimated by maximum likelihood. The log likelihood is given by

$$l(y; \Lambda, \Psi) = -\frac{N}{2} \log |\Lambda\Lambda^T + \Psi| - \frac{1}{2} tr[S(\Lambda\Lambda^T + \Psi)^{-1}],$$

where $S = \sum_i (y - \mu_Y)(y - \mu_Y)^T$. Score functions cannot be solved directly, so iterative procedures are often used. Alternately the likelihood can be maximised using the expectation maximisation algorithm (EM; Dempster et al., 1977).

Extensions of latent variable models can expand the use of this very powerful technique further. Spatial and temporal factor analysis (Wang & Wall, 2003) are latent factors that are smooth in space or time. These can model unobserved variables which are assumed to be spatially or temporally smooth, as would be the case for

environmental variables like temperature and rainfall. They can be used as a simple way to model spatial and temporal correlation between variables, that arises as a result of several separate processes (factors), some of which may act over longer and others over shorter spatial or temporal scales. Another extension of latent variable models is structural equation models (Sánchez et al., 2005). These models are related to confirmatory factor analysis: a model is hypothesised for how observed and latent variables are interrelated, and model summary measures are used to compare competing models, or assess the viability of a proposed model. A further extension is sparse factor analysis (Meng et al., 2014), where the matrix of loadings is encouraged to be sparse.

As discussed in Section 2.3, latent variable models have recently been used in multivariate hierarchical Bayesian models in ecology, both to induce sparsity, and to find patterns in covariance between species. In Chapter 5 we describe a novel algorithm for applying any covariance modelling algorithm designed for Gaussian data to multivariate abundance and other discrete multivariate data.

4.1.2 Gaussian graphical models

Gaussian graphical models (Banerjee et al., 2006; Meinshausen & Bühlmann, 2006; Yuan & Lin, 2007; Friedman et al., 2008; Rothman et al., 2008) describe conditional independence relationships between Gaussian variables. For a P-variate Gaussian random variable $Z \sim N(0, \Sigma)$, variables j and j' are conditionally independent, given the rest, if the (j, j') element of $\Theta = \Sigma^{-1}$ is zero (Baba et al., 2004). To estimate graphical models, modern methods impose an L_1 penalty on the elements of Θ to encourage sparsity. The problem is thus reduced to maximising a penalised log likelihood, namely

$$\log|\Theta| - \mathrm{tr}(S\Theta) - \lambda||\Theta||_1, \tag{4.1}$$

where tr() denotes the trace and $||\Theta||_1$ is the L_1 norm (sum of absolute values). Fast algorithms for solving this constrained optimisation iteratively employ the coordinate descent algorithm (Friedman et al., 2008). The subgradient Equation for (4.1)

is

$$\Theta^{-1} - S - \lambda\Gamma = 0,$$

where $\Gamma_{q,r} = \text{sign}(\Theta_{q,r})$ if $\Theta_{q,r} \neq 0$ and $\Gamma_{q,r} \in [-1, 1]$ if $\Theta_{q,r} = 0$. The resulting graph can be interpreted as the pattern of direct and indirect relationships among variables, or more formally conditional dependence.

Graphical models are not commonly implemented for ecological data, with some recent exceptions. Harris (2015) describes a Markov model for binary data based on the well known Ising graphical model (Ravikumar et al., 2010), while Morueta-Holme et al. (2016) implements graphical modelling on transformed abundances, both with the intention of inferring species interactions.

4.2 Covariance modelling of discrete data

While covariance modelling of Gaussian data is well understood, covariance modelling of discrete data remains challenging. Models for Gaussian data have been extended to the discrete data setting, though generally separately for each covariance modelling paradigm and a particular discrete distribution. In addition many of these do not allow joint modelling of a response to predictors with covariance modelling.

Flexible tools for graphical and latent factor modelling of discrete and mixed data can be achieved by using non-parametric marginal distributions (Carvalho et al., 2008; Liu et al., 2009; Gruhl et al., 2013; Murray et al., 2013; Abegaz & Wit, 2014; Fan et al., 2016; Abegaz & Wit, 2015; Guo et al., 2015). As interest is on the covariance structure, these treat marginal distributions as nuisance parameters. and by design do not include predictors. An alternative formulation for covariance modelling without covariates is item response theory, which allows latent variable modelling for binary and multinomial data (Hambleton, 1991).

For joint modelling of predictors and covariance of counts and categorical outcomes, generalised latent variable models (Skrondal & Rabe-Hesketh, 2004; Holst & Budtz-Jørgensen, 2013) provide a flexible covariance modelling method. These models

are able to fit generalised linear mixed models and structural equation models in a combined framework. However, these cannot be used to carry out other forms of covariance modelling, and are restricted to the Poisson distribution for counts, which generally fits poorly for multivariate abundance data (ver Hoef & Boveng, 2007). Resorting to a Poisson distribution with latent variables for overdispersed counts can induce unintended artefacts (Section 2.4).

It is possible to build graphical models for discrete data by extending Gaussian graphical models to other members of the exponential family (Banerjee et al., 2006; Lee & Hastie, 2015). These models include the well known Ising models for binary data. The Ising model can be fit using approximations to the intractable normalising constant (Banerjee et al., 2006; Höfling & Tibshirani, 2009; Ravikumar et al., 2010), but is difficult to fit in large dimensions. Additionally, for other marginal distributions, including the Poisson for counts, these extensions place restrictions on the direction of conditional relationships between variables. To overcome this limitation for counts, node wise graphical models have been proposed (Allen & Liu, 2013), however these are local in nature and do not estimate a global model of dependence, making them inefficient.

There are a number of quite flexible Bayesian methods for covariance modelling of discrete and mixed data, these include Bayesian graphical models (Pitt et al., 2006; Smith & Khaled, 2012) as well as latent variable models (Murray et al., 2013). Bayesian copula graphical models are implemented in Dauwels et al. (2013) and Guo et al. (2015) via a conditional EM algorithm.

4.3 Covariance modelling of multivariate abundance data

Latent variable models and graphical models make quite different and interesting assumptions about the structure of the data. Graphical models assume most variables are conditionally independent of one another, and hence the precision (inverse covariance) matrix is sparse. For multivariate abundance data, where variables are species, we can interpret these conditional dependence relationships as species inter-

actions. This definition of "species interactions" applied here concerns two species that interacting with one another not in terms of the species abundance being correlated or co-occurring, but being correlated after accounting for the effect of all other species and environment filtering. The output graph gives us information about how correlations between species are controlled by common interactions with intermediate species. Importantly, these models explain all correlations between species as interactions between them. On the other hand, latent variable models propose correlation between species as a result of unobserved or missing covariates. In ecology, these are often interpreted as unmeasured environmental variables, which drive dependence between species. Model based ordination plots (Ovaskainen et al., 2016; Hui, 2016), obtained by plotting latent factors and scores, can be used to visualise the drivers of site and species differences. Both these structures are plausible explanations for correlation between species, and it is of interest to ecologists to explore and test such relationships.

The power of covariance modelling techniques is the ability to find patterns in dependence between variables, and in particular those not explained by known covariates. To model multivariate abundance data, where environmental covariates have a strong effect on how species are related, the effects of these have to be accounted for if covariance modelling is to give us useful information on species interactions. In Chapter 5 we propose a novel method to apply any covariance modelling method for Gaussian data to discrete data.

In addition to finding patterns in covariance between species, covariance models can substantially reduce the number of parameters in the covariance matrix, while making only mild assumptions on the structure of the matrix. This property allows us to conduct inference on multivariate abundance data, where there are generally few observation relative to the number of species. In Chapter 7 we propose a method for likelihood based marginal inference for discrete data using copulas and covariance modelling.

Chapter 5

Covariance modelling of discrete data with copulas

The main appeal of copula models is their flexibility, as the covariance structure can be modelled separately from marginal models. In this chapter will extend the flexibility of Gaussian copulas further, by allowing existing algorithms for covariance modelling of Gaussian data to be utilised for estimating the covariance structure when the responses are non-Gaussian. This will allow us to combine any collection of marginal distributions, and any covariance modelling algorithm designed for Gaussian data.

As we discussed in Chapter 4, covariance modelling techniques, like factor analysis and graphical models, can find patterns in covariance for multivariate data, as well as allow the building of parsimonious multivariate models, when there are few observations relative to the number of variables. These are both desirable properties for multivariate models of abundance data in ecology, where we often have a small sample size relative to the number of variables.

In this chapter we will describe a flexible algorithm which can combine any set of discrete (or continuous) response distributions, and any covariance modelling algorithm designed for Gaussian data, in a Gaussian copula framework. We then demonstrate the use of this algorithm with factor analytic and graphical covariance models.

5.1 Model formulation

We model discrete response matrix y as a copula with marginal parameter vectors (β, ψ), and correlation matrix R_θ, parameterised by a set of variables θ. The likelihood (from Equation 3.4), is given by

$$L(y|\beta,\psi,\theta) = \prod_{i=1}^{N} \int_{A_i} |R_\theta|^{-1/2} \exp\left(-\frac{1}{2} z_i^T (R_\theta^{-1} - I) z_i\right) du_i, \tag{5.1}$$

where $z_{ij} = \Phi^{-1}(u_{ij})$ and $A_i = \cap_j \left[(F_{ij}(y_{ij}^-|\beta_j,\psi_j), F_{ij}(y_{ij}|\beta_j,\psi_j)\right]$. Here F_{ij} are the marginal distributions with marginal parameters β_j and ψ_j.

5.2 Estimation

We implement a type of Monte Carlo expectation maximisation (MCEM; Wei & Tanner, 1990) algorithm to estimate this integral. We chose an algorithm which is easy to implement, and allows the flexibility we desire. We will start by defining the the MCEM algorithm, Gaussian score equation and Dunn-Smyth residuals (Dunn & Smyth, 1996).

Definition 1 (Monte Carlo Expectation Maximisation)**.** The expectation maximisation algorithm (EM; Dempster et al., 1977) is a method to maximise the likelihood function in the presence of missing data z. This is done iteratively; in the E-Step one calculates the Q function,

$$Q(\theta, \hat{\theta}^{(m)}) = \int_{z_i} f(z|y; \hat{\theta}^{(m)}) \log f(z;\theta) dz,$$

which is the expectation of the log likelihood with respect to the conditional predictive distribution $f(z|y; R_{\hat{\theta}^{(m)}})$, under the current value of the model parameters $\hat{\theta}^{(m)}$ at the mth iteration. The Q function is then maximised in the M-Step to find the new value of the model parameters,

$$\hat{\theta}^{(m+1)} = \arg\max_\theta Q(\theta, \hat{\theta}^{(m)}).$$

These steps are repeated iteratively until convergence. When the Q function is not available in closed form, a Monte Carlo estimate of the required expectation

can be used instead. This is the Monte Carlo Expectation Maximisation algorithm (MCEM; Wei & Tanner, 1990). The Q function is replaced by

$$\tilde{Q}(\theta, \hat{\theta}^{(m)}) = \frac{1}{K} \sum_{k=1}^{K} \log f(z_k; R_{\hat{\theta}^{(m)}}),$$

in the E-Step, where z_k, $k = 1, ..., K$ are drawn from $f(z|y; \hat{\theta}^{(m)})$.

Definition 2 (Gaussian score equation for covariance parameters)**.**

$$\frac{\partial l(\mathbf{z}; R_\theta)}{\partial \theta} = \sum_{i=1}^{N} \frac{\partial \log N_P(z_i^k; R_\theta)}{\partial \theta}. \tag{5.2}$$

Definition 3 (Dunn-Smyth residuals)**.** Dunn-Smyth residuals are a useful diagnostic tool for generalised linear modelling (Dunn & Smyth, 1996), but are used here as a device for numerical approximation of the integrand in Equation (5.1). Let u_{ij} be a uniform random variable. A Dunn-Smyth residual can be defined as

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

where $F_{ij}(y_{ij}^-) = \lim_{x \to y_{ij}^-} F_{ij}(x)$. The distribution of these residuals, given the marginal distributions, is a truncated multivariate normal with identity covariance matrix.

$$g(\boldsymbol{z_i}) = \frac{\prod_{j=1}^{P} \phi(z_{ij}^k)}{\prod_{j=1}^{P} f_{ij}(y_{ij})} I_{z \in A_i},$$

where $A_i = \cap_j \left[(F_{ij}(y_{ij}^-|\beta_j, \psi_j), F_{ij}(y_{ij}|\beta_j, \psi_j)\right]$.

This distribution has positive probability only in the region of integration of the copula likelihood (Equation (5.1)), making it a candidate for importance sampling to estimate this integral. Importance sampling schemes using these and similar constructs appear by other names in Heinen & Rengifo (2007), Nikoloulopoulos (2013b) and others, where the resulting approximations are maximised numerically, as discussed in Chapter 3.4. Instead of implementing a numerical optimisation scheme, we note in the following theorem that the copula score function, when approximated by importance sampling with Dunn-Smyth residuals, can be rewritten as a weighted sum of Gaussian score equations. This allows us to maximise the

copula likelihood using the very algorithms used to carry out covariance modelling of Gaussian data.

Lemma 1. The discrete Gaussian copula likelihood can be approximated by importance sampling with K sets of Dunn-Smyth residuals

$$\begin{aligned} L_i(\mathbf{y}_i; R_\theta) &= \int_{A_i} |R_\theta|^{-1/2} \exp\left(-\frac{1}{2} z_i^T (R_\theta^{-1} - I) z_i\right) du_i \\ &\approx \prod_{j=1}^{P} f_{ij}(y_{ij}) \sum_{k=1}^{K} c(\mathbf{z}^k; R_\theta), \end{aligned} \tag{5.3}$$

where $N_P(\cdot; R)$ is a multivariate Gaussian density with zero mean, unit variance and correlation matrix R, $c(\cdot; R_\theta) = N_P(\mathbf{z}_i^k; R_\theta)/\prod_{j=1}^{P} \phi(z_{ij}^k)$ is the Gaussian copula density with correlation matrix R_θ, $f_{ij}(\cdot)$ is the marginal distribution of variable j and observation i. The proof in Appendix A follows by importance sampling arguments.

We now demonstrate the link between the Gaussian score equation, and the Gaussian copula score.

Theorem 1. An estimate of the derivative of the Gaussian copula likelihood with discrete margins (with respect to covariance parameters) can be written as a weighted sum of derivatives of the multivariate Gaussian distribution.

Proof: Differentiating the log likelihood approximation from Lemma 1 we have

$$\begin{aligned} \frac{\partial l(\mathbf{y}; R_\theta)}{\partial \theta} &= \sum_{i=1}^{N} \frac{1}{\sum_{k=1}^{K} c(\mathbf{z}_i^k; R_\theta)} \sum_{k=1}^{K} \frac{\partial c(\mathbf{z}_i^k; R_\theta)}{\partial \theta} \\ &= \sum_{i=1}^{N} \frac{1}{\sum_{k=1}^{K} c(\mathbf{z}_i^k; R_\theta)} \sum_{k=1}^{K} \frac{\partial c(\mathbf{z}_i^k; R_\theta)}{\partial \log c(\mathbf{z}_i^k; R_\theta)} \frac{\partial \log c(\mathbf{z}_i^k; R_\theta)}{\partial \theta} \\ &= \sum_{i=1}^{N} \frac{1}{\sum_{k=1}^{K} c(\mathbf{z}_i^k; R_\theta)} \sum_{k=1}^{K} c(\mathbf{z}_i^k; R_\theta) \frac{\partial \log c(\mathbf{z}_i^k; R_\theta)}{\partial \theta} \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{c(\mathbf{z}_i^k; R_\theta)}{\sum_{k=1}^{K} c(\mathbf{z}_i^k; R_\theta)} \frac{\partial \log c(\mathbf{z}_i^k; R_\theta)}{\partial \theta} \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} w_{ik}(R_\theta) \frac{\partial \log \phi_P(\mathbf{z}_i^k; R_\theta)}{\partial \theta}, \end{aligned} \tag{5.4}$$

where $w_{ik} := c(\mathbf{z}_i^k; R_\theta)/(\sum_{k=1}^K c(\mathbf{z}_i^k; R_\theta))$

Covariance modelling algorithms, like those which estimate a factor analysis or structured covariance matrices, maximise the Gaussian likelihood by design, or equivalently solve the Gaussian score equations. By writing the copula score equation as a weighted sum of the Gaussian scores, we are able to utilise these algorithms with a weighted set of the Dunn-Smyth residuals. As the weights w_{ik} are a function of the parameters to be estimated, these must be iteratively updated, and so we arrive at the algorithm below.

5.3 Algorithm

To carry out covariance modelling on discrete data with a Gaussian copula, we iteratively implement the covariance modelling algorithm designed for Gaussian data on a weighted set of Dunn-Smyth residuals.

Algorithm 1 Covariance modelling for discrete data

For data y and covariates X

1. Estimate $F_{ij}(\cdot; X_i)$ using a univariate modelling algorithm (*e.g.* `glm`)
2. For $k = 1, \cdots, K$, simulate Dunn-Smyth residuals

$$z_{ijk} = \Phi^{-1}\{\hat{F}_{ij}(y_{ijk} - 1) + u_{ijk}\hat{f}_{ij}(y_{ij})\}$$

3. Initialise $w_{ik}^{(0)} \propto 1$ and write $(z, w^{(m)})$ for the set of Dunn-Smyth residuals and weights
4. For $m = 1, 2, \cdots$, until convergence
   a Apply covariance modelling algorithm to weighted residuals $(z, w^{(m-1)})$ to obtain $\hat{\theta}^{(m)}$
   b Recalculate weights $w_{ik}^{(m)} \propto c(z_{ik}; R_{\hat{\theta}^{(m)}})$ from Theorem 1

Note: As most covariance modelling algorithms use the sample covariance matrix as a sufficient statistic, we can in practice use the weighted correlation matrix of Dunn-Smyth residuals $R_w^{(m)} = \frac{1}{N}\sum_{i=1}^N \sum_{k=1}^K w_{ik}(R_{\hat{\theta}^{(m-1)}}) z_i^k (z_i^k)^T$ as a sufficient statistic in step 4a.

This algorithm has two estimation steps. Firstly, marginal parameters (β, ψ) are estimated assuming independence, as with independence estimating equations (Liang & Zeger, 1986). Secondly, these estimates $(\hat{\beta}, \hat{\psi})$ are plugged into the copula likelihood $L(y|\beta, \psi, \theta)$ (Equation 5.3). The resulting plug-in likelihood $L(y|\hat{\beta}, \hat{\psi}, \theta)$ is maximised for covariance parameters θ using an iterative procedure, which can be understood as a MCEM algorithm (McLachlan & Krishnan, 1997) where the sample for the E-step is achieved by reweighting the residuals, and the M-Step is the covariance modelling algorithm, see Appendix A.3 for proof.

The integral in equation 5.1 can be estimated numerically in other ways, including quadrature (see for example: Song et al., 2009). However, the derivation above leads most naturally to the MCEM algorithm described. This algorithm can easily incorporate existing covariance modelling algorithms designed for Gaussian data, without the need for alteration.

The derivation of the approximate likelihood in derivative in Theorem 1 leads very naturally to the MCEM algorithm described. Alternative methods to estimate the required integrals are possible, including various quadrature methods.

We obtain consistent estimates for all model parameters (see Appendix A.2 for proof). This algorithm extends the flexibility of Gaussian copulas to implement any covariance modelling framework designed for Gaussian data to discrete data.

5.4 Application to covariance modelling methods

Algorithm 1 can be implemented with covariance models estimated by maximum likelihood as well as penalised likelihood. We will demonstrate this with two examples, graphical modelling for penalised likelihood and factor analysis for maximum likelihood.

5.4.1 Application to graphical models

Modern implementations of graphical modelling for Gaussian data optimise a penalised likelihood with a lasso penalty (Banerjee et al., 2006). Though this is not

a maximum likelihood algorithm, as required by Theorem 1, we will show that Algorithm 5.3 can nevertheless be used to carry out graphical modelling of discrete data.

We begin by applying the relevant likelihood penalty to the approximate log likelihood in Theorem 1. Let $\Theta = R^{-1}$ be the precision matrix. The penalised log likelihood estimate can be written;

$$l^{\lambda}(\mathbf{y};\Theta) = \left[\sum_{i=1}^{N}\sum_{j=1}^{P}\log(f_{ij}(y_{ij}))\right] + \sum_{i=1}^{N}\log\left[\sum_{k=1}^{K}c(\mathbf{z}_i^k;\Theta)\right] - \lambda||\Theta||_1,$$

from Lemma 1. To find the maximiser of this function we write

$$\begin{aligned}
0 = \frac{\partial l^{\lambda}(\mathbf{y};\Theta)}{\partial\Theta} &= \sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik}(\Theta)\left[\frac{\partial\log\phi_P(\mathbf{z}_i^k;\Theta)}{\partial\Theta}\right] - \lambda\Gamma \\
0 &= \sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik}(\Theta)\left[\Theta - \mathbf{z}_i^k\mathbf{z}\text{'}_i^k\right] - \lambda\Gamma \\
0 &= \hat{\Theta} - \left[\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik}(\hat{\Theta})\mathbf{z}_i^k\mathbf{z}\text{'}_i^k\right] - \lambda\Gamma, \qquad (5.5)
\end{aligned}$$

where $\Gamma_{q,r} = \text{sign}(\Theta_{q,r})$ if $\Theta_{q,r} \neq 0$ and $\Gamma_{q,r} \in [-1,1]$ if $\Theta_{q,r} = 0$. Equation (5.5) is analogous to the Equation (4.1) solved by Gaussian graphical modelling algorithms like the graphical lasso (Friedman et al., 2008), with the sample covariance matrix replaced by a weighted covariance matrix with weights w_{ik}. We can therefore solve Equation (5.5) iteratively using the graphical lasso algorithm together with Algorithm 1.

5.4.2 Application to factor analysis

As discussed in Section 4.1.1, factor analysis is solved by maximum likelihood, either by numerically solving the score equations or with the EM algorithm. The numerical algorithms are implemented to solve the following score equations (Everitt, 1984).

$$\begin{aligned}
0 &= \text{diag}\left(\Sigma^{-1}\left[\frac{1}{N}\sum_{i=1}^{N} y_i y\text{'}_i\right]\Sigma^{-1} - \text{diag}(\Sigma^{-1})\right) \\
0 &= \Sigma^{-1}\left[\frac{1}{N}\sum_{i=1}^{N} y_i y\text{'}_i\right]\Sigma^{-1}\Lambda - \Sigma^{-1}\Lambda.
\end{aligned}$$

Now looking at the copula likelihood estimate we have

$$l(\mathbf{y};\Lambda,\Psi) = \left[\sum_{i=1}^{N}\sum_{j=1}^{P}\log(f_{ij}(y_{ij}))\right] + \sum_{i=1}^{N}\log\left[\sum_{k=1}^{K}c(\mathbf{z}_i^k;R)\right],$$

where $R = \Lambda\Lambda^T + \Psi$. We differentiate with respect to both Λ and Ψ to obtain

$$\begin{aligned}
0 &= \frac{\partial l(\mathbf{y};R)}{\partial\Psi} = \sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik}(R)\left[\frac{\partial\log\phi_P(\mathbf{z}_i^k;R)}{\partial\Psi}\right] \\
0 &= \sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik}(R)\left[\mathrm{diag}(R^{-1}z_i^k z'^k_i R^{-1} - \mathrm{diag}(R^{-1}))\right] \\
0 &= \mathrm{diag}\left(R^{-1}\left[\frac{1}{N}\sum_i\sum_{k=1}^{K} w_{ik}(R)z_i^k z'^k_i\right]R^{-1} - \mathrm{diag}(R^{-1})\right). \qquad (5.6)
\end{aligned}$$

Similarly the derivatives with respect to Ψ give us score equations

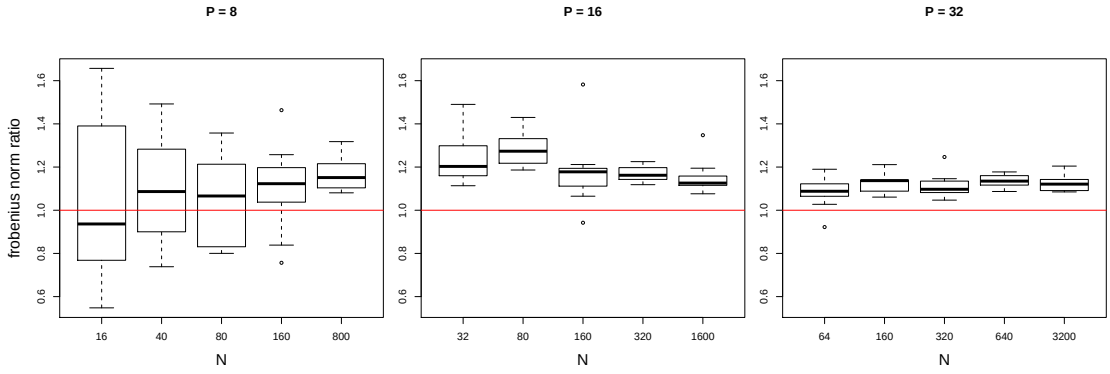
$$0 = R^{-1}\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik}(R)z_i^k z'^k_i\right]R^{-1}\Lambda - R^{-1}\Lambda.$$

Comparing these to the score equations for factor analysis for Gaussian data, it is clear we can replace the data covariance matrix with a weighted covariance matrix of Dunn-Smyth residuals with weights w_{ik}, and use standard factor analysis algorithms, iteratively updating the weights, to model the correlation matrix R.

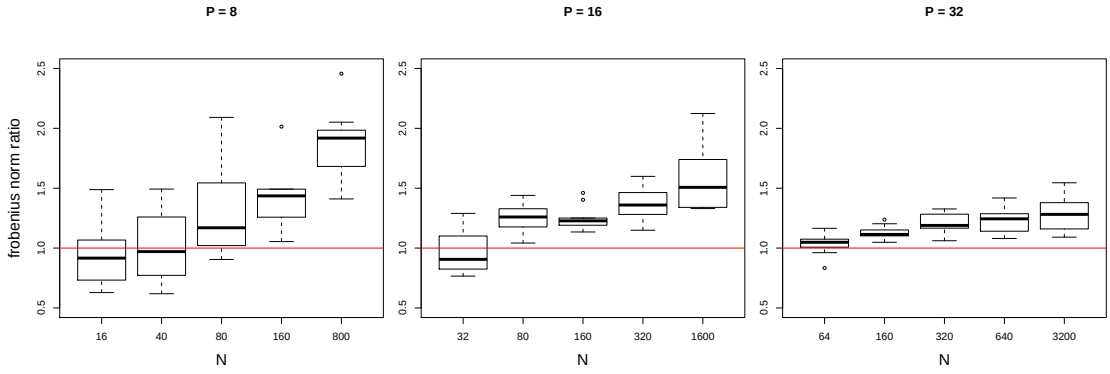
5.5 Simulation results

5.5.1 Factor analysis: Binary data

We compare Algorithm 1 to two alternative strategies for factor analysis of discrete data. We generate a one factor binomial model for binary data with probit link using the `lava.tobit` (Holst, 2012) package. This package is able to simulate and estimate a probit regression with latent factors. In this special case, the Gaussian copula is equivalent to a hierarchical model (Nikoloulopoulos, 2013b), and thus can be fitted using software for hierarchical latent variable modelling, like `lava.tobit`. Additionally we will compare to a naive procedure in which we carry out a factor analysis on Pearson residuals or one set of Dunn-Smyth residuals from a binomial generalised linear model. Both these sets of residuals should be approximately normally distributed marginally, and so a factor analysis algorithm can be applied di-



(a)



(b)

Figure 5.1: Comparison of Algorithm 1 to the `lava.tobit` package (b) and factor analysis of Pearson residuals (a). Values above the red line indicate the copula model is performing better. Algorithm 1 generally outperforms these alternatives, especially as dimension increases.

rectly to these residuals for an approximate solution. For further simulation details see Appendix B.1.1.

We will measure the performance of factor analysis models by the Frobenius norm (as in Ledoit & Wolf (2003)) of the difference of estimated and true covariance matrices. Figure 5.1a shows that Algorithm 1 generally outperforms the naive application of factor analysis algorithms to Pearson residuals. One set of Dunn-Smyth residuals performs similarly to Pearson residuals, and we do not include these results. Figure 5.1b shows that with as few as 50 sets of Dunn-Smyth residuals Algorithm 1 generally outperforms the `lava.tobit` package in terms of accuracy.

5.5.2 Graphical model: Count data

For graphical modelling we simulate data from a Gaussian copula with Poisson marginal distributions, and a chosen graphical structure. We then measure how well our model, and others, are able to discover the graphical structure. We generated and estimated graphical structures and data using the `huge` package (Zhao et al., 2012) in `R`, which simulates and estimates Gaussian graphical models. Graphical modelling works best for sparse matrices, so the graphs we generate have a 70% probability for conditional independence for any pair of variables. For model selection we use the StARS criterion (Liu et al., 2010) which chooses the model with the most stable graphical structure across sub samples. We then compare Algorithm 1 to the local Poisson model (Allen & Liu, 2013) as well as a naive application of a graphical modelling algorithm to Pearson and one set of Dunn-Smyth residuals. For further simulation details see Appendix B.1.2.

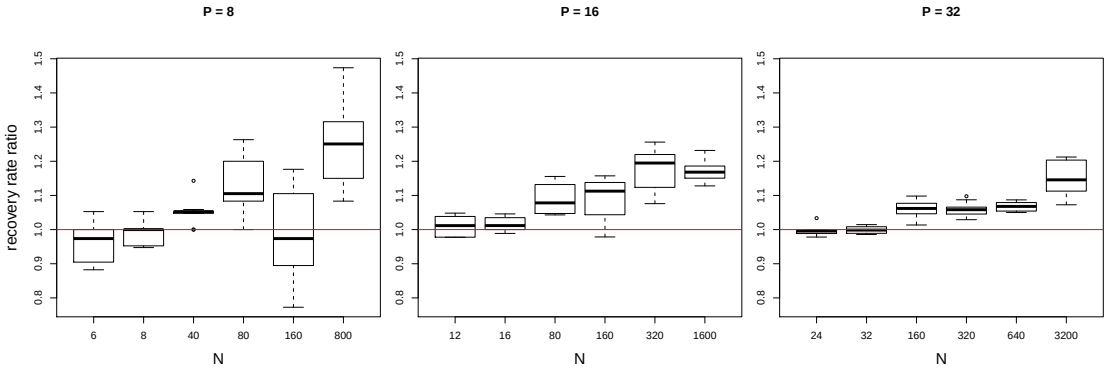
We measure the performance of the graphical modelling algorithms as the proportion of correctly identified conditional dependence relationships. Figure 5.2a shows that Algorithm 1 generally outperforms the naive application of graphical modelling algorithms to Pearson residuals, one set of Dunn-Smyth residuals performs similarly to Pearson residuals and is not shown. Algorithm 1 also generally outperforms the local Poisson model (Figure 5.2b), particularly as dimension (P) increases.

Poisson distributed counts were simulated for easy comparison to the local Poisson model, but note our model can easily be extended to modelling overdispersed counts by using a negative binomial regression in the marginal model, as below.

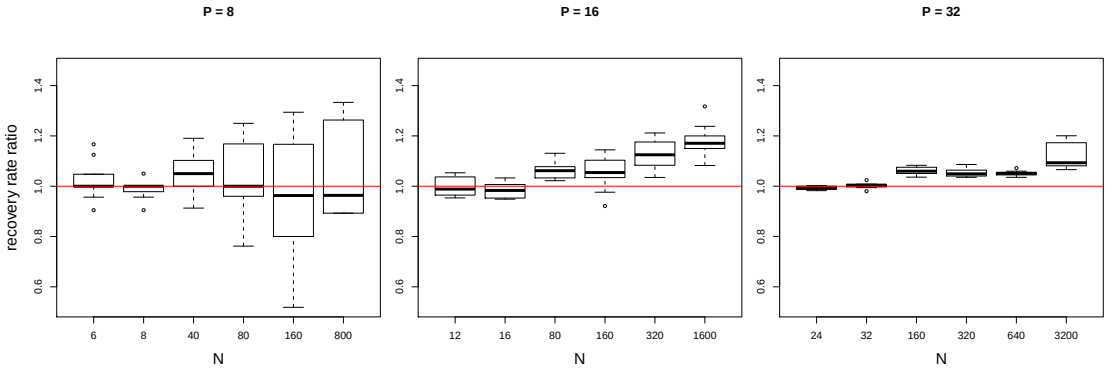
5.6 Practical Application

5.6.1 Count data: Hunting spider data

We will demonstrate our method on the spider dataset (Section 1.1.2) which contains counts of the number of hunting spiders caught in traps for 12 species taken from 28 sites, modelled as a function of environmental variables. For these data we fit marginal negative binomial generalised linear models with all the available



(a)



(b)

Figure 5.2: Comparison of Algorithm 1 to graphical modelling of Pearson residuals (a) and the local Poisson model (b; Allen & Liu, 2013). Values above the red line indicate the copula model has a higher recovery rate, (*i.e.* proportion of correctly identified conditional dependence relationships) and is therefore performing better. Algorithm 1 generally outperforms these alternatives, especially as dimension increases.

environmental covariates using the `mvabund` package (Wang et al., 2012b), which also contains these data. We then use Algorithm 1 to estimate the graph of conditional independences using the graphical lasso implemented in the `glasso` package (Friedman et al., 2008). We also carry out a factor analysis using the `factanal` function in `base R` (R Core Team, 2014).

Figure 5.3 has the output of a factor analysis and graphical model before controlling for covariates (left) and after (right). The first row are factor scores in a two factor model, the second row are loadings for each species, and the third row are the graphs obtained from a graphical model.

In our ecological example we study covariance relationships before ad after accounting for correlation due to the environmental covariates. With variables representing different species, we could interpret the graphical model as a model of species interactions, and attempt to identify which species interact directly with one another, and which are correlated due to their interaction with common species. The factor analysis of these data highlights latent factors which drive correlations among species, which may be unmodelled environmental variables.

We have coded the plots of scores (Figure 5.3a-b) according to the covariates in Figure 1.3, the presence of bare sand (filled for present and unfilled for absent), and the presence of fallen leaves (blue triangle for present and red square for absent). We observe clustering in Figure 5.3a according to both variables. Sites with bare sand and no fallen leaves (filled squares) load negatively on both factors (bottom left of Figure 5.3a), while sites with fallen leaves and no bare sand (unfilled triangles) load positively on factor 2 and negatively on factor 1 (top left of Figure 5.3a). No patterns are visible after controlling for these covariates (Figure 5.3b).

Additionally there are patterns among species in Figure 5.3c-f. For example, species *Pardlugu (Pardosa lugubris)* has negative interactions with both *Alopacce (Alopecosa accentuata)* and *Pardmont (Pardosa monticola)*, who interact positively with one another, before controlling for covariates (Figure 5.3e). However these negative interactions are absent after controlling for covariates (Figure 5.3f). Looking at Figure 1.3, we can see that *Pardlugu* has decreased abundance in the presence of bare sand

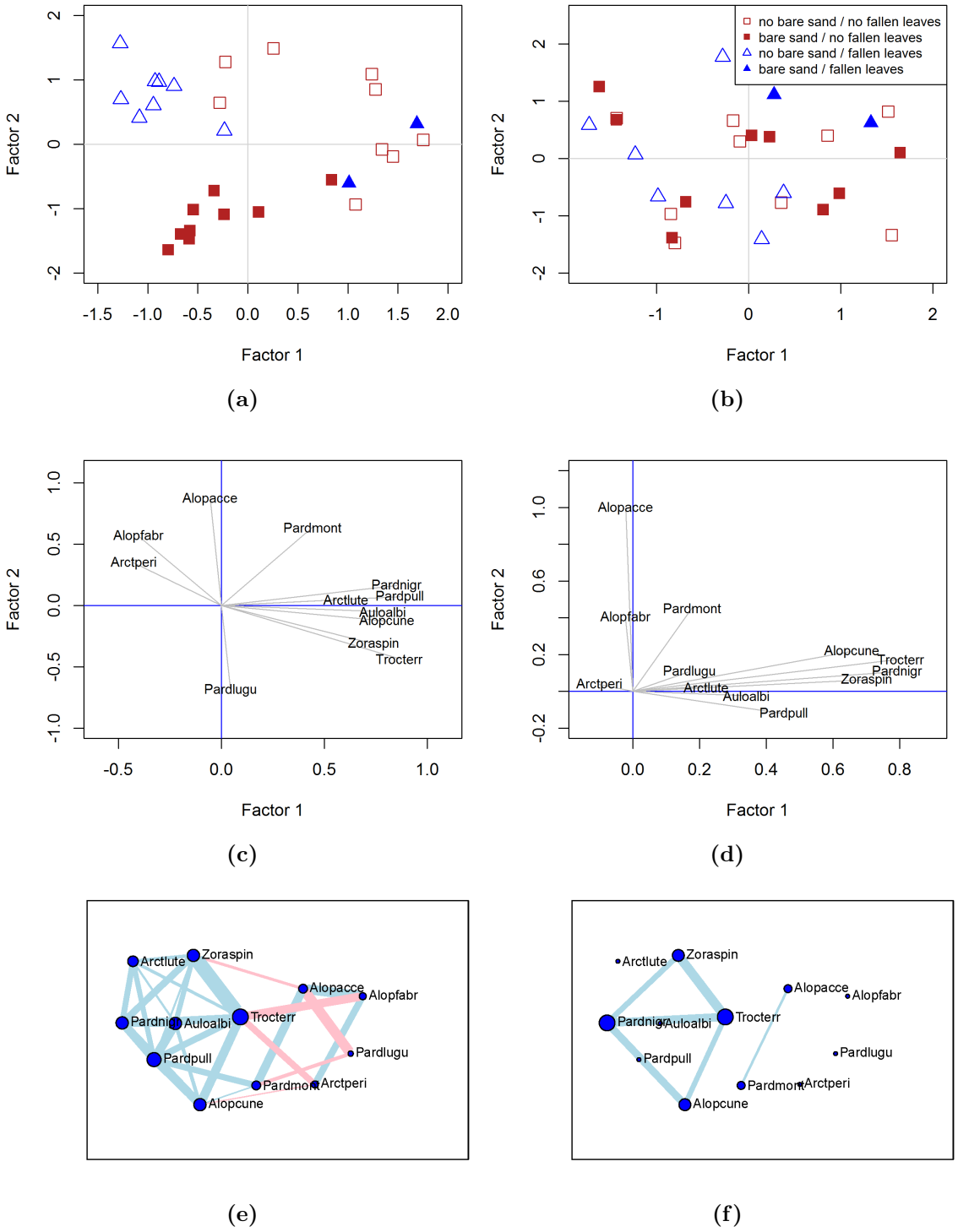


Figure 5.3: Results of covariance models before (left) and after (right) controlling for covariates. Factor scores (a and b), factor loadings (c and d) and graphs (e and f). We observe clustering of sites in terms of these covariates in (a), while after controlling for covariates, these patterns are absent in (b). Figure (c) to (f) also show responses to covariates. *Pardlugu* has negative interactions with both *Alopacce* and *Pardmont*, who interact positively, before controlling for covariates (e), which are absent after controlling for covariates (f). *Pardlugu* has the opposite response to covariates to *Alopacce* and *Pardmont* (Figure 1.3), so these interactions may be explained by the covariates.

and increased abundance in the presence of fallen leaves, while both *Alopacce* and *Pardmont* have the opposite response to these covariates. The difference between Figure 5.3e and f may therefore be due to these interactions being explained by these covariates. These patterns are consistent with the factor loading plots Figure 5.3c-d. In Figure 5.3c, before controlling for covariates, *Pardlugu* loads negatively on factor 2, while *Alopacce* and *Pardmont* load positively on factor 2 (Figure 5.3c). After controlling for covariates, *Pardlugu* has very small loading on both factors, indicating it is not strongly correlated to any of the species (Figure 5.3d). The interaction between *Alopacce* and *Pardmont* remains positive even after controlling for covariates (Figure 5.3f), and they both load positively on factor 2 (Figure 5.3d) in the full model.

5.7 Discussion

We have developed a general algorithm for covariance modelling of discrete data. It can combine any likelihood based covariance modelling procedure designed for Gaussian data, with any set of marginal distributions, and is simple and flexible. The algorithm we present does not place restrictions on the sign of covariance parameters, nor is it restricted to one or a small class of covariance models. It is fully flexible in terms of both the marginal distributions and covariance parameters, and only assumes the covariance structure of the latent variable is that of a multivariate Gaussian, and marginal distributions are correctly specified.

Simulation results show our method is not only more general than alternative proposals, but also has advantages in performance. For graphical modelling of counts, our model outperforms the local Poisson model (Allen & Liu, 2013), and has the further advantage that it can additionally accommodate covariates and overdispersion. For factor analysis of binary data, our method also outperforms the `lava.tobit` package on R. An alternative approach we also considered was to perform covariance modelling on a single set of residuals from univariate models, but this seemed to lose considerable efficiency.

The method described has many advantages, most notably the flexibility not offered

by other methods, as well as advantages in statistical performance. It is however less computationally efficient than many of the methods we compared to, including all the methods which use one set of residuals, as well as the `lava.tobit` package, which uses composite likelihood (Lindsay, 1988). That being said, the computational burden of our method is not overly large, with graphical analysis of 10 species and 1000 sites taking less than 1 second, 40 species and 1000 sites less than 90 seconds, and 1000 species with 1000 sites taking less than 3 hours. These times including model selection to select sparseness along a path of 100 shrinkage values. The code has not been optimised with C++ or Fortran, and this is planned for future research.

In addition, inference for many alternate covariance models for discrete data is conditional (see section 2.3), while copula inference is marginal, which changes the interpretation of model parameters (see section 2.4). These two types of models are therefore used to answer different questions.

We demonstrate our method with two well known covariance modelling frameworks, but it is simple to substitute other (possibly penalised) likelihood-based covariance modelling algorithms for Gaussian data. Covariance modelling of Gaussian data is a fast moving area of research (Section 4), with new methods often being developed. Our algorithm can accommodate these new covariance modelling methods as soon as they become available, and allow them to be used with discrete data. There is also no reason that all the marginal distributions need to be from the same family, nor do they need to all be discrete. All combinations of covariance modelling algorithms and marginal distributions can be accommodated.

Chapter 6

Species interactions in New Zealand native forests

The introduction, explanation of conditional independence (Section 6.1), visualising multivariate abundance data (Section 6.2), as well as Methods (Section 6.3.1) and data analysis was carried out by GCP with input from supervisors FKCH and DIW. This chapter was completed in collaboration with Joanna M. Buswell[3], who supplied the data, and Angela T. Moles[1] and Fiona J. Thomson[2], who assisted with interpretation of results from an ecological perspective and suggest phrasing for ecological aspects of the Results and Discussion sections.

1. School of Biological, Earth and Environmental Sciences, UNSW Australia, NSW 2052, Australia

2. Landcare Research, Lincoln, 7640, New Zealand

3. Ministry for the Environment, Wellington, New Zealand

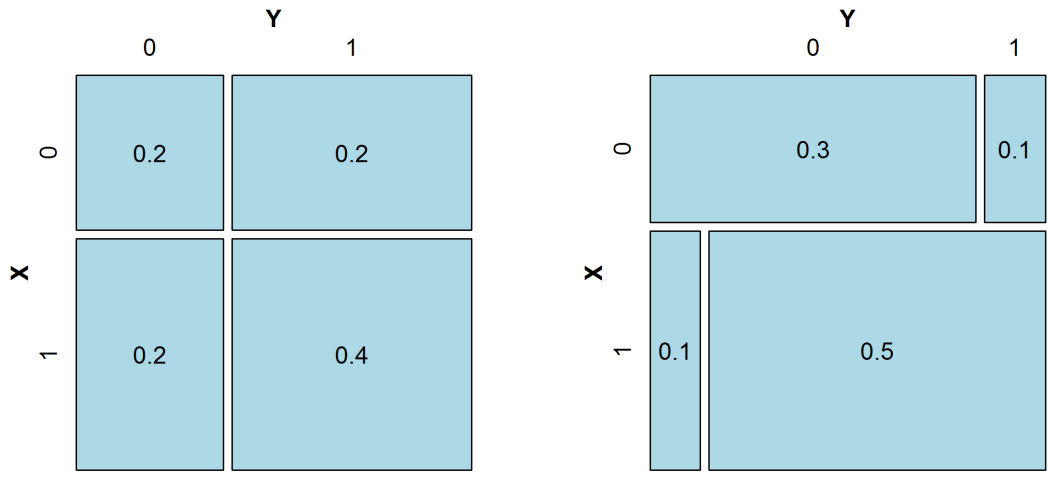
How species of plants and animals interact with one another is an important question of interest to community ecologists. This question can be investigated in a number of ways. One method is to directly observe species interacting (for example pollination or predation), and build models based on these observed interaction networks (Jordano et al., 2003; Wells & O'Hara, 2013). Alternately one can measure

the abundance of a small number of species over time, and estimate how species abundance influences future abundance of other species (Brown et al., 2001; Carrara et al., 2015). Both these methods require data to be collected specifically for the purpose of studying interactions, and are limited to studying a small number of species. In addition, interaction networks can be inferred from proxies like functional traits, geographical distributions and phylogenies (*e.g.* Morales-Castilla et al., 2015), in the absence of empirical data.

It is also possible to extract information on species interactions from routinely collected co-occurrence data (multivariate abundance data). For this type of data, interactions have been studied using null models (Gotelli & Ulrich, 2010; Strong Jr et al., 2014) and more recently hierarchical models (Section 2.3; Pollock et al., 2014; Ovaskainen et al., 2016). In both cases, species interactions are defined in terms of correlations. Species are thought to interact if they are correlated, possibly after accounting for known covariates. There are however several reasons species might be correlated. These include a joint response to missing covariates, or a common interaction with other species in the community. The first of these can be studied with latent variable models (Section 4.1.1), which can be interpreted as modelling missing covariates. Factor loadings from latent variable models provide information about how species respond to missing covariates.

To distinguish between species which interact directly, and those which are correlated due to shared interactions with other species, we need to investigate conditional dependence relationships, which can be studied with graphical models (Section 4.1.2). Graphical models have not been widely employed in ecology, with some exceptions (Harris, 2016; Morueta-Holme et al., 2016). They are however becoming popular in other biological fields such as neuroscience (Huang et al., 2010; Allen et al., 2012), and gene expression studies (Schultz et al., 2012; Allen & Liu, 2013).

In this chapter we will analyse ordinal data in the form of percent cover categorised into six categories (Table 1.4). Graphical modelling of ordinal data can be carried out with Bayesian models (Dobra & Lenkoski, 2011; Mohammadi & Wit, 2015). In a likelihood framework graphical models for ordinal data are generally fitted semi-parametrically (Liu et al., 2009; Guo et al., 2015), with non-parametric marginal



(a) X and Y independent.

(b) X and Y dependent.

Figure 6.1:
Mosaic plot example when X and Y are independent (a) and dependent (b).

distributions, such that the marginal distributions are modelled non-parametrically without reference to any covariates.

In Chapter 5 we described a method for applying covariance modelling techniques, including graphical models, to discrete data, which can be presence/absence, biomass, count and ordinal. In this chapter we will demonstrate conditional dependence with a simple example. We will then describe how graphical models can differentiate between correlations between species, and conditional dependence (which we interpret as species interactions). We then contrast graphical modelling with other methods of visualising high dimensional data in ecology. Finally we use graphical models to investigate species interactions in New Zealand native forests.

6.1 Conditional independence

The concept of conditional independence is best illustrated with an example. For two binary variables, their dependence can be displayed in a mosaic plot. When variables are independent, mosaic plots have parallel rectangles both vertically and horizontally (Figure 6.1a), while dependent variables do not (Figure 6.1b).

For our example, we generate presence/absence data for three species, A, B, and C, such that species B and C interact with A, but not with one another. We do this by first generating species A. We then generate species B with probabilities of presence conditional on the presence or absence of species A, but irrespective of the presence of species C, and vice versa. The relationship between species A and B, and A and C are displayed in Figure 6.2a-b respectively. For example when A is absent (A=0), species B is more likely to be absent, while species C is more likely to be present. As both B and C depend on A, they are not generally independent (Figure 6.2c). The information in these plots is pairwise between species, and is related to the information we obtain from correlations. For this example, all three species are correlated (Figure 6.2d).

We simulated this example such that species B and C do not interact directly, their presence depends only on the presence of A. Figure 6.2 demonstrated that pairwise metrics, such as correlation, cannot extract these interactions, as interactions are conditional. In Figure 6.3 we go one step further to show the joint probabilities of each pair of variables conditional on the third. Figure 6.3 (a) and (b) exhibit conditional dependence. However, in Figure 6.3c, we see that conditional on A, species B and C are independent (the mosaic sub-plots for A=0 and A=1 have parallel rectangles). We can now observe that while species B and C are correlated, and dependent, they are independent conditional on A.

The purpose of graphical models is to find these conditional relationships. These model precision matrices, which specify conditional independence patterns, rather than correlation matrices. As in the above example, it is possible to have a dense matrix of correlations (all correlations between A, B and C are non-zero), but a sparse precision matrix (the conditional dependence between B and C is 0).

6.2 Visualising multivariate abundance data

We present graphical models as a powerful technique for visualising multivariate abundance data, which shows information on species interactions. In ecology the most common methods for visualising multivariate data are ordination techniques.

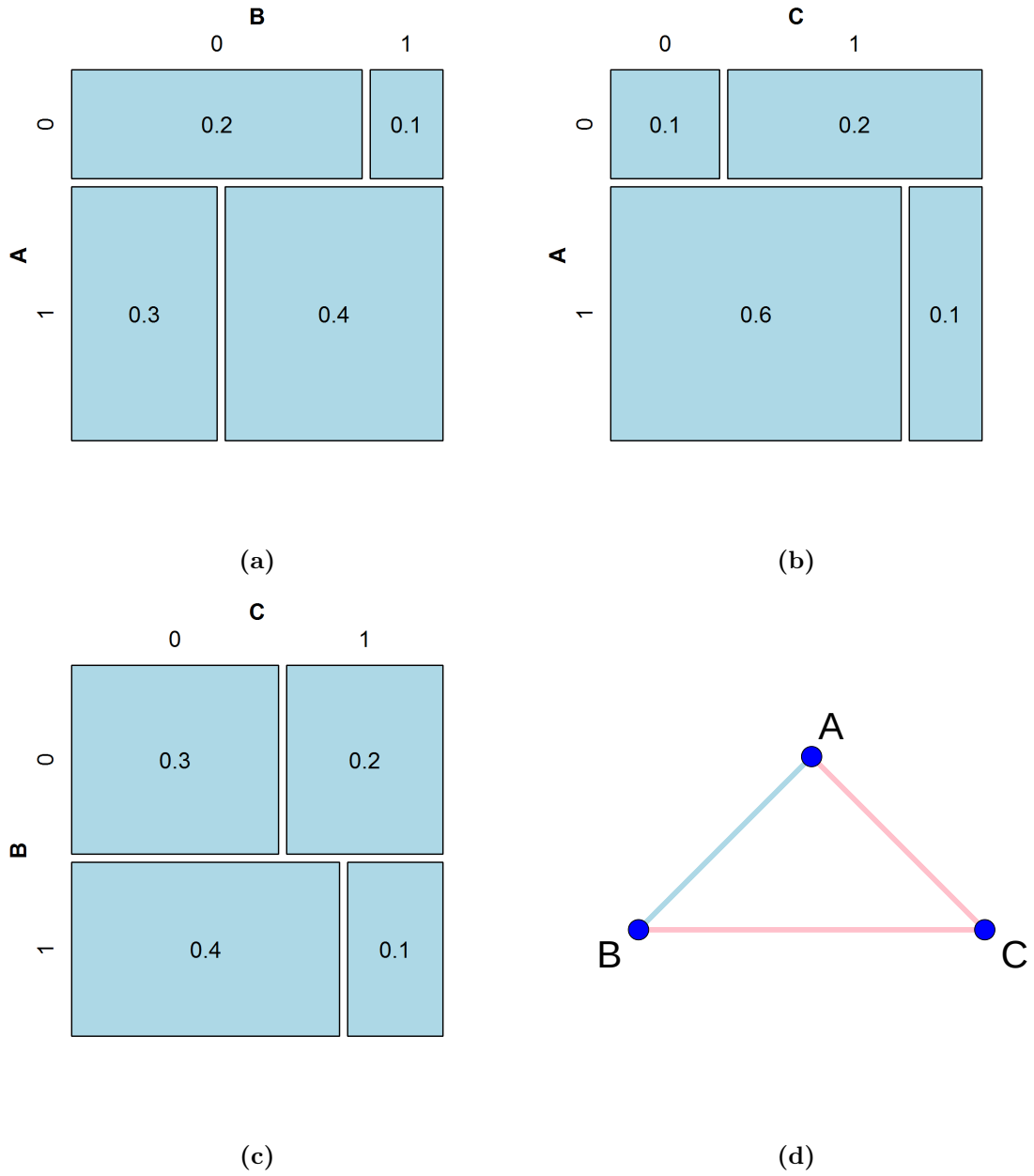
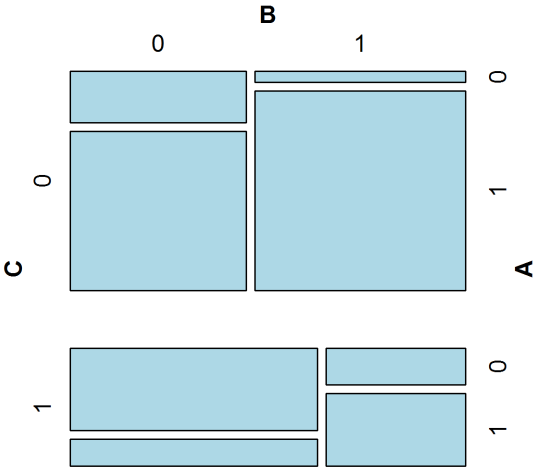
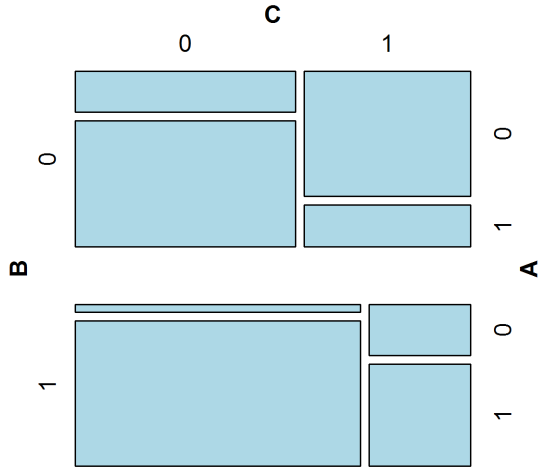


Figure 6.2: Dependence between species (a) A and B, (b) A and C, and (c) B and C. No pair of species is independent.

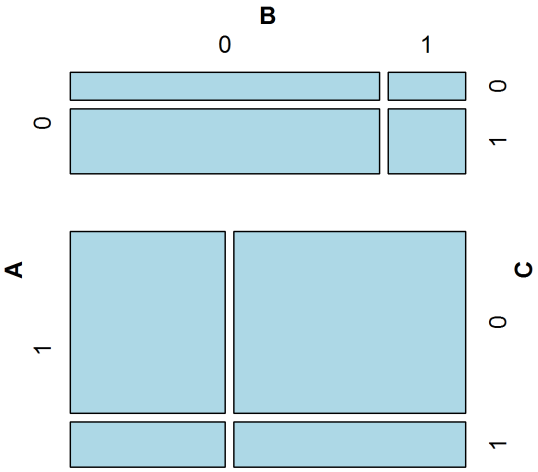
(d) Correlation graph of species A, B, and C showing correlations (lines) between all pairs of species, with blue lines for positive correlation and red for negative.



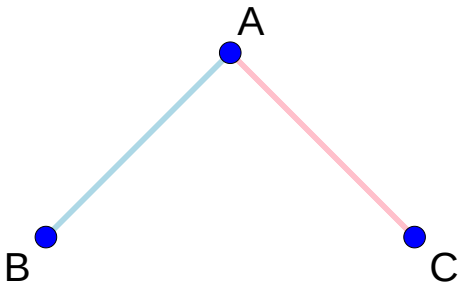
(a) A and B conditional on C.



(b) A and C conditional on B.



(c) B and C conditional on A.



(d) Interaction.

Figure 6.3: Conditional dependence for species (a) A and B conditional on C, (b) A and C conditional on B and (c) B and C conditional on A. No conditional independence observed in (a) or (b). Plot (c) has parallel rectangles in the sub-plots of B and C conditional on A being either one or zero, so B and C are conditionally independent given A.
(d) Interaction graph of species A, B, and C showing independence (no line) between B and C conditional on A, positive dependence (blue line) between A and B conditional on C, and negative dependence (red line) between A and C conditional on B.

Ordination is a generic name given to methods which reduce multivariate data from many response variables, to just two, in order to display the data on a scatter plot. In ecology, by far the most common method of ordination is non-metric multidimensional scaling (nMDS; Kruskal, 1964). This is an algorithmic method which rearranges points in two dimensions, such that the ordering of pairwise distances between points best matches the ordering of pairwise dissimilarities between observations (sites). These dissimilarities are commonly calculated using a metric such as Bray-Curtis (Bray & Curtis, 1957) on (possibly transformed) data.

Recently, model based methods for ordination, using latent variables, have become increasingly popular for visualising data (Hui et al., 2015a; Ovaskainen et al., 2016). Model based ordination gives roughly the same interpretation as an nMDS plot, and serves as a method to reduce many response variables to a two dimensional scatter plot. Sites similar to one another in terms of species composition or relative abundance tend to be highly correlated, and so cluster together. However, model based ordination has several advantages over distance based techniques such as nMDS. They appropriately handle the properties of multivariate abundance data, including the mean-variance relationship, which can confound trends in location with trends in dispersion if not accurately modelled (Warton et al., 2012). In addition these models are based on likelihoods, and inherit the desirable properties of likelihood inference, including model selection methods and good predictive capacity (Hui et al., 2015a). In addition the axes of the resulting ordination plots (or latent variables) can be interpreted as missing covariates in the model. Graphical models are related to latent variable models, as both model covariance between variables (Chapter 4). The main distinction between them is in the information they illicit about correlations. Graphical models display information on conditional dependence, and hence species interactions, which no other visualisation method currently used in ecology is able to do.

While we believe this provides a powerful new way to visually investigate multivariate abundance data, it does come with some strong assumptions. Graphical models assume that most species do not interact with one another, and give misleading results if this is not the case. For multivariate abundance data this amounts to an

assumption that most species interact directly with only a few others. This does not imply that most species are not correlated; a sparse precision matrix can induce a dense correlation matrix. In addition, graphical models, like other penalised likelihood methods (Hastie et al., 2015) are only reliable as a measure of conditional independence when there is a lot of data relative to the number of variables, which is often not the case for multivariate abundance data. Also, for discrete data, the conditional independence relationships obtained from copula graphical models (Section 5.4.1) are on a latent scale, and do not necessarily imply conditional independence of the discrete data. They do however imply near conditional independence for ordinal variables with a large number of categories (Abegaz & Wit, 2014). For these reasons we present graphical models as primarily an exploratory tool.

The main output of graphical modelling is a 'graph' (we will refer to this as an interaction graph for clarity), which for P species can be represented as a P by P matrix $\boldsymbol{G}$ of species interactions, with $\boldsymbol{G}_{j,k} = 0$ if species j and k do not interact, and $\boldsymbol{G}_{j,k} = 1$ if they do. It is common to plot interaction graphs with vertices being species, and lines only between species which interact (as in Figure 6.3d). We can also obtain the sign of the interaction (positive or negative) as well as a relative measure of the strength of interactions. Modern graphical modelling techniques use penalised likelihood (Tibshirani, 1996), with a penalty parameter λ which controls the total number of interactions. We can obtain a sparse graph with very few interactions, or a dense graph with many interactions, from the same data. Using the method described in Chapter 5, we can obtain an interaction graph for the raw data, or we can first account for covariates and obtain a residual graph. This is conceptually similar to residual ordination (Ovaskainen et al., 2016; Hui, 2016), sometimes referred to as partial ordinations. Residual interaction graphs do not contain species interactions that are explained by a shared response to covariates included in the model.

6.3 Data analysis

6.3.1 Methods

We implement graphical modelling using Algorithm 1 from Section 5.3 with a cumulative link (Agresti, 2010) marginal model to the New Zealand native forest cover data (Section 1.1.3). These data contain observations of cover (in ordinal classes) for 1311 species at 964 sites. We included slope and altitude as covariates. The final graph is chosen by BIC (Section 7.2.3). For plotting we leave out species which did not interact with any other species from all graphs. We employ the Fruchterman Reingold algorithm (Fruchterman & Reingold, 1991) to position nodes in two dimensions. This algorithm attempts to position edges in two dimensional space such that the edge lengths are equal and there are as few crossings as possible.

6.3.2 Results

We have analysed 1311 species at 964 sites. Of these, 142 were found to have interactions with other species. Due to the large number of interacting species, the interaction graph of all species (Figure 6.4) is not visually very informative, nevertheless we can see some important patterns. Most New Zealand forest species are positively associated with one another (blue lines), with the two ends of this gradient of positive association being negatively associated (as shown by the u-shaped group of positive associations dominating). The dominant species at one end of the spectrum tend to be associated with silver beech forest (*Lophozonia menziesii* (NOTMEN), *Raukaua simplex* (RAUSIM), *Myrsine divaricata* (MYRDIV), *Blechnum procerum* (BLEPRO), *Coprosma foetidissima* (COPFOE), and *Notogrammitis billardierei* (GRABIL)), while the other end of the gradient is dominated by species associated with the early-mid stages of regeneration of disturbed lowland forest (*Cyathea dealbata* (CYADEA), *Melicytus ramiflorus* (MELRAM), *Knightia excelsa* (KNIEXC) and *Uncinia uncinata* (UNCUNC)).

For more informative plots, we can zoom in on certain subsets of species, to better examine interactions between individual species. These subset interaction graphs

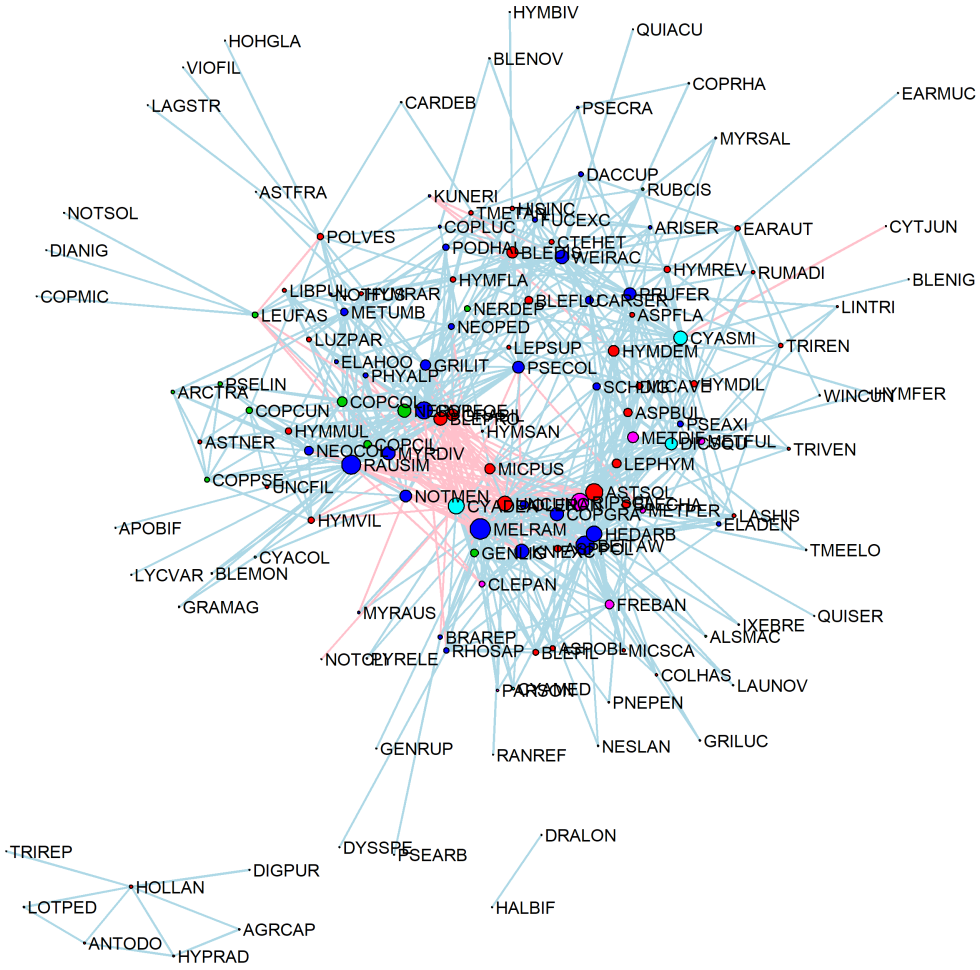


Figure 6.4: Interaction graph for all species after controlling for covariates (slope and altitude). There are a range of positive (blue) and negative (red) interactions between species. Species which do not interact directly have no lines joining them. The colour of the dot corresponds to species type; herbs (red), shrubs (green), trees (blue), vines (magenta) and tree ferns (cyan). Species at one end of the spectrum (NOTMEN, RAUSIM, MYRDIV, BLEPRO, COPFOE, and GRABIL) are associated with silver beech forest, while the other end of the gradient is dominated by species associated with the early-mid stages of regeneration of disturbed lowland forest (CYADEA, MELRAM, KNIEXC, and UNCUNC).

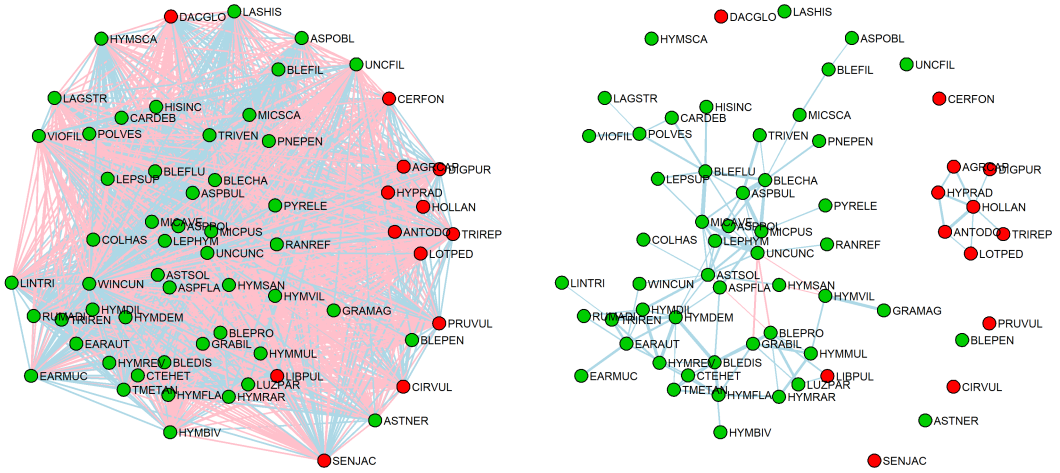


Figure 6.5: Herb interactions; the interaction graph (right) is much more informative than the correlations induced by these interactions (left). The interaction graph clearly shows a group of exotic (red) herbaceous (TRIREP, DIGPUR, HOLLAN, AGRCAP, LOTPED, ANTODO, HYPRAD)) that are positively associated with one another, and not associated with any native (green) species. The correlation graph is too dense to interpret.

display interactions between the species after accounting for covariates and interactions with all other species. To contrast interaction graphs with correlation, in Figure 6.5 we have plotted all the correlations in (a), and the interaction graph in (b), for all herb species in the data. This is similar to Figure 6.2d and 6.3d, where the interactions between species A and C, and species A and B, induce correlation between species B and C. The interaction graph (Figure 6.5b) shows a clear distinction between native herbaceous species (green points), and exotic herbaceous species (red points). The group of exotic herbaceous species on the right of the figure (*Trillium repens* (TRIREP), *Digitalis purpurea* (DIGPUR), *Holcus lanatus* (HOLLAN), *Agrostis capillaris* (AGRCAP), *Lotus pedunculatus* (LOTPED), *Anthoxanthum odoratum* (ANTODO) and *Hypochaeris radicata* (HYPRAD)) are positively associated with one another, and not associated with any native species. In addition *Hymenophyllum sanguinolentum* (HYMSAN) and *Hymenophyllum villosum* (HYMVIL) have a negative interaction even though they can co-occur (Brownsey & Perrie, 2014). No patterns are distinguishable on the correlation graph.

Lastly we plot interaction graphs before and after controlling for covariates. Looking

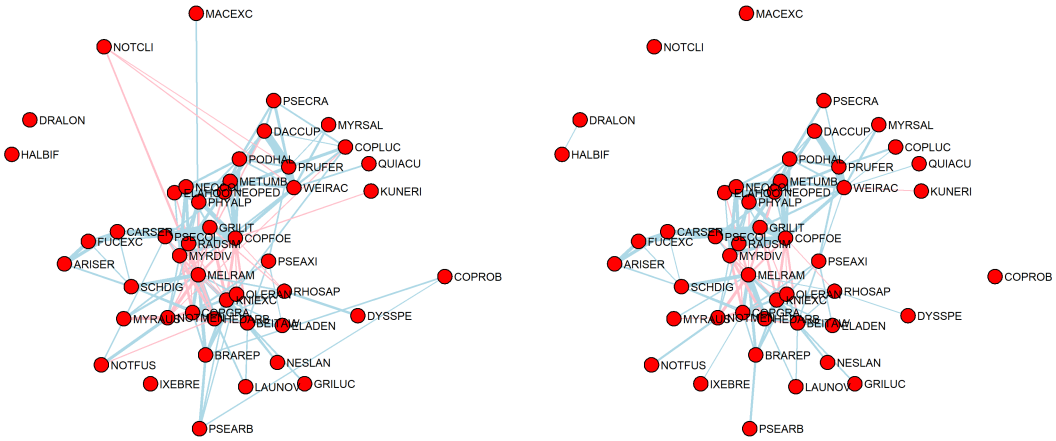


Figure 6.6: Interaction graph of trees before (a) and after (b) controlling for covariates . Negative interactions between NOTCLI and other species are not present after controlling for covariates (slope and altitude).

at these graphs can suggest which interactions are explained by the covariates modelled. For example negative interactions between *Fuscospora cliffortioides* (NOTCLI) and other species are present on the graph not controlling for any covariates (Figure 6.6a) but absent after controlling for altitude and slope (Figure 6.6b).

6.4 Discussion

In general, interaction graphs are quite informative, and much more so than the associated graph of correlations (Figure 6.5). In Figure 6.5b we observe a clear distinction between native herbaceous species (green points), and exotic herbaceous species (red points). This pattern is consistent with the fact that these exotic species are not shade tolerant, and thus differ from many of the native herbaceous species (a group which includes a range of understorey ferns) in not being able to survive in forest understoreys. The associated plot of correlations (Figure 6.5a) is too complex to distinguish any patterns.

As this system is well studied, we would expect the interaction graphs to be consistent with known interactions between species. For example, the negative association between beech forest and broadleaf forest (Figure 6.4) was consistent with our initial

expectation that there would be a separation between the two main forest types of New Zealand. This prediction was based on the observation that podocarp-broadleaf forest and beech forests tend not to co-occur, but are not obviously separated by geography or climate. In addition, the fact that most interactions are positive is consistent with Figure 1.5, which indicates most correlations are positive for these species.

In Figure 6.6 we observed negative interactions between *Fuscospora cliffortioides* (NOTCLI) and other species, which are present before controlling for covariates, but absent after controlling for covariates (including altitude). This is understood to have happened because the lack of co-occurrence is explained by environment. *Fuscospora cliffortioides* occurs in montane and subalpine forest that tends to occur at altitudes between 400 m - 1380 m above see level (Wiser et al., 2011), whereas *Prumnopitys ferruginea* (PRUFER), *Weinmannia racemosa* (WEIRAC) and *Raukaua simplex* (RAUSIM) occur in lower altitude forest types, that range from sea-level up to 700 m in the south island and up to 1100 m in the north island (Wiser et al., 2011).

Some of the interactions we found were surprising, and can be used to generate hypotheses for further investigation. In Figure 6.4, we observed most beech forest species interacting positively, and broadleaf-podocarp interacting positively, with negative interactions separating the two forest types. However, this separation was not supported for the three species in *Fuscospora*, the other genus of southern beech present in New Zealand (*Fuscospora solandri* (NOTSOL), *F. cliffortioides* (NOTCLI), *F. fusca* (NOTFUS). The three *Fuscospora* species fell in different parts of the graph rather than clustering together or with *Lophozonia menziesii*, and were not associated with many other species (either negatively or positively).

The negative interaction between *Hymenophyllum sanguinolentum* (HYMSAN) and *Hymenophyllum villosum* (HYMVIL) (Figure 6.5) is also surprising. These two species can co-occur (Brownsey & Perrie, 2014) but are often confused for each other. This negative interaction may therefore be an artefact of misclassification rather than a true negative interaction. Field ecologists may identify one or the other of these species, and assume everything similar in a plot is the same herb.

The result would be that the species would rarely be recorded to co-occur.

For the New Zealand native forest data, graphical models have been able to confirm known patterns of interaction, as well as generate hypotheses based on surprising interactions. Overall, graphical models have allowed the visualisation of a large and complex dataset, to better understand the ecology of species interaction. Hence this technique of fitting graphical models to discrete data *via* Gaussian copulas is a potentially valuable exploratory tool for multivariate analysis in ecology.

Chapter 7

Multivariate inference for discrete data with Gaussian copulas

In this chapter, we propose a likelihood based method of marginal inference for multivariate abundance data, and other correlated discrete data. Inferring the relationship between environmental and experimental variables and the community of species is at least partly the aim of all our motivating datasets (Section 1.1), and most multivariate abundance data. As discussed in Chapter 2, conditional inference for multivariate abundance data is implemented with hierarchical models, generally estimated with Bayesian methods (Section 2.3: Walker & Jackson, 2011; Ovaskainen et al., 2016; Warton et al., 2015a, and others). Common approaches to marginal inference for multivariate discrete data are generalised estimating equations (GEEs; Liang & Zeger, 1986; Zeger & Liang, 1986) and copula models (Sklar, 1959), although inference for multivariate abundance data in particular is almost always implemented with GEEs (Section 2.2).

To carry out marginal inference with GEEs, Wald and score tests only require a covariance matrix for model parameters, which can be estimated even when there are many species relative to sample size (Warton, 2011). They are therefore the most common way to conduct marginal inference in this context. In addition, score tests only require the calculation of the null model, giving them a substantial computational advantage over both Wald and likelihood ratio tests. These tests however are

known to suffer from poor power for data with properties very commonly seen in multivariate abundance data. The Wald test statistic is known to lose power when conditional means are near a boundary, such as when species are rare, and hence overall species means are low. The score test statistic (and related information criteria), on the other hand, can have poor power when the distribution of covariates is skewed, for example when sampling is unbalanced, another common characteristic of these data. We demonstrate that a likelihood ratio test and likelihood based information criteria do not have these disadvantages, and are better suited to these data.

In order to conduct a likelihood ratio test, we must first define and estimate a multivariate likelihood. GEEs only define a likelihood when variables are assumed to be independent, which has a detrimental impact on power (Section 7.1.1). In Chapter 5, we describe a method for covariance modelling of discrete correlated data using copulas. In this chapter we will demonstrate how discrete copulas, with covariance models, can be used to carry out inference about community-environment associations for correlated discrete data. We will use a Gaussian copula likelihood for multivariate abundance data to conduct likelihood ratio tests and model selection with standard information criteria. We demonstrate desirable power properties of these statistics in the context of analysing multivariate abundance data.

7.1 Hypothesis testing

7.1.1 Modelling covariance *vs.* assuming independence

It is straightforward to construct a likelihood ratio test for multivariate data by assuming independence between variables. Warton (2011), for example, investigates the performance of this test, relative to a test statistic that incorporates covariance between variables. When the covariate effect is along the dominant eigenvector, the direction in which the data are most variable, then incorporating covariance actually leads to poorer power relative to assuming independence. When the effect is orthogonal to the dominant eigenvector, tests which estimate dependence tend to have better power.

Figure 7.1a demonstrates this effect. For a bivariate Gaussian random variable, we have plotted likelihood ratio statistics derived from models assuming independence (left) and models estimating the correlation (right). The likelihood ratio statistic at the maximum likelihood estimator (MLE; round dot) is equal to one, by definition. As we move away from the MLE, the likelihood ratio statistic decreases, and consequently the power of tests increases, and indicated by the intensity of colour. When assuming independence, the power increases uniformly in all directions away from the MLE. By contrast, when correlations are estimated, the power increases more quickly orthogonal to the main eigenvector, to reflect the direction in which data are least variable. As a consequence, the relative power of these two statistics depends critically on the direction of the covariate effect.

Let us assume the covariate is a treatment, where the effect of treatment can be in the same direction as the covariance of the response (square) or orthogonal to it (triangle). For example, with two positively correlated species, a treatment effect in the direction of the covariance implies treatment has the same effect for both species, either increasing or decreasing both abundances, while a treatment effect orthogonal to the covariance implies that the treatment increases abundance for one species, but decreases abundance for the other. When testing for a treatment effect along the main eigenvector (square), then a test assuming independence is more powerful (square is in darker region on left then right). On the other hand, when the covariate effect is orthogonal to the main eigenvector (triangle), a likelihood ratio test which assumes covariance has better power (triangle is in darker region in on the right than left)

So interestingly, assuming independence when variables are correlated can lead to superior power in particular circumstances. Of course, we generally don't know the direction of the effect in practice, and so would prefer a test which is more powerful under most circumstances, and this is illustrated in Figure 7.1b. In these plots, red indicates regions where estimating correlation is more powerful, while in blue regions it is better to assume independence. For uncorrelated data, both tests are asymptotically equivalent, but as correlation increases ($\rho = 0.5$ on the left and $\rho = 0.9$ on the right), we can see red regions, where the correlation is estimated,

are dominant, implying which a test statistic which estimates correlation tends to be more powerful in general.

7.1.2 Power comparison of Wald, score, and likelihood ratio statistics

We start by defining the Wald, score and likelihood ratio tests statistics. The Wald test statistic is defined as

$$W = (\hat{\theta} - \theta_0)^T \hat{\Sigma}(\theta)^{-1} (\hat{\theta} - \theta_0),$$

where $\hat{\Sigma}(\theta)$ is an estimate of the covariance of θ, and θ_0 is the value of θ under the null hypothesis. We use the sandwich estimator given in Equation (2.4). The score test statistic is given by

$$S = u(\theta)^T \hat{\Sigma}(\theta)^{-1} u(\theta),$$

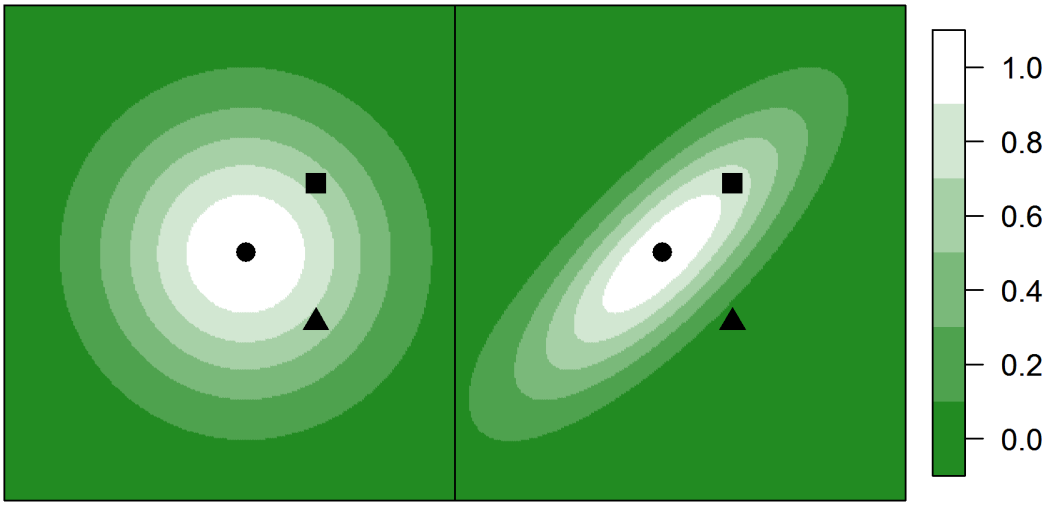
where $u(\theta_0) = \frac{\partial \log L(\theta)}{\partial \theta}|_{\theta_0}$ is the score. The likelihood ratio statistic is

$$LR = 2 \log L(\hat{\theta}) - 2 \log L(\theta_0).$$

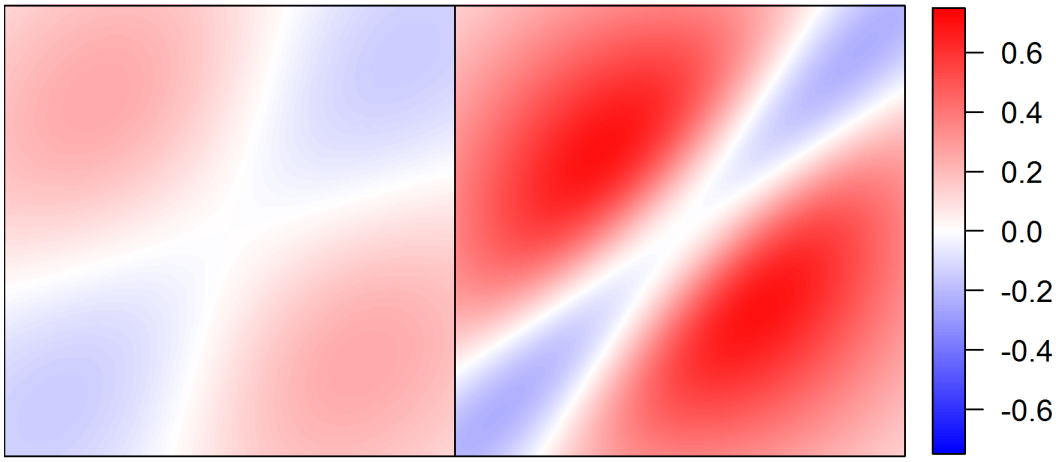
All three are asymptotically χ^2_d distributed when the null model is correct, where d is the difference in the number of parameters between the two models.

As discussed in Section 2.2, GEEs can be used for marginal inference on correlated discrete data. Specifically, hypothesis tests that only require an estimate of covariance, such as the Wald and score tests, can be carried out in the GEE framework. Wald, score and likelihood ratio statistic are asymptomatically equivalent (Rao, 2009), however finite sample properties differ depending on the data and model. Several power studies have shown no clear superior test asymptotically and by simulation (Sutradhar & Bartlett, 1993; Lemonte & Ferrari, 2012; Dobek et al., 2015).

Importantly though, is that the Wald and score statistics have some undesirable properties. The Wald statistic does not increase monotonically when the observed value of the test statistic is on the boundary of the parameter space (Væth, 1985) and as effect size increases (Hauck Jr & Donner, 1977). Both Wald and score tests also present extreme behaviour for unbalanced experimental designs (*i.e.*, either



(a)



(b)

Figure 7.1:

a) Likelihood ratio statistic for bivariate Gaussian assuming (left) independence and (right) correlated likelihood. As we move away from the MLE (round dot) the power of the test statistic increases, and indicated by the intensity of colour. When testing for a covariate effect along the main eigenvector (square), the test assuming independence is more powerful (square is in darker region on the left than right). On the other hand, when the covariate effect is orthogonal to the main eigenvector (triangle) the likelihood which estimates covariance has better power (triangle is in darker region on right than left).

b) Relative power when we assume independence *vs.* estimate correlation when correlation is i) 0.5 and ii) 0.9. Red indicates estimating correlation leads to better power. As correlation increases (left to right), the region where estimating correlation leads to better power dominates.

very low or very high power, Warton, 2008), as will be illustrated in the example below. Multivariate abundance data often have these properties, including unbalanced sampling designs, and more generally skewed predictors. Very small means are also common due to most species being rarely observed. This tends to place test statistics on or near the boundary of the parameter space. In these circumstances likelihood ratio tests can be expected to behave more consistently and have better power in general.

Consider a negative binomial regression with one binary predictor *e.g.* a treatment factor with two levels.

$$\begin{aligned} Y &\sim NegBin(\lambda_i, \phi) \\ \log(\lambda_i) &= X_i^T \beta \\ X^T &= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & \dots & 1 \\ 0 & 0 & \dots & 0 & 1 & \dots & 1 \end{bmatrix}, \end{aligned}$$

with N observations in total, n_0 in the control and $n_1 = N - n_0$ in the treatment group. Under this simple model, it is straightforward to derive the Wald, score and likelihood ratio statistics to test for an effect of treatment, and consider their behaviour. The statistics are given by

$$\begin{aligned} W(\beta_1) &= \frac{(\log \bar{y_1} - \log \bar{y_0})^2 n_0 n_1 \bar{y_0} \bar{y_1}}{\bar{y_0} n_0 + \bar{y_1} n_1 + \phi \bar{y_0} \bar{y_1} N} \\ S(\beta_1) &= \frac{(\bar{y_1} - \bar{y_0})^2 n_0 n_1}{N \bar{y}(1 + \phi \bar{y})} \\ LR(\beta_1) &= 2 \sum_{g=1}^{2} n_g \bar{y_g} \log \left(\frac{\bar{y_g}}{\bar{y}} \right) - 2 \sum_{g=1}^{2} n_g (\bar{y_g} + \phi^{-1}) \log \left(\frac{1 + \phi \bar{y_g}}{1 + \phi \bar{y}} \right), \end{aligned}$$

where $\bar{y_g}$ is the observed mean of group g, and ϕ is the overdispersion parameter.

In Figure 7.2a we plot the observed test statistic against the proportion of observations in the less variable group, with observed means for the null and alternate group set to $\bar{y}_0 = 0.1$ and $\bar{y}_1 = 1$ respectively, overdispersion held at $\phi = 2$ and a sample size of $N = 10$. If the variance of the two groups was not relevant to power, then we would expect this graph to be symmetric around 0.5. However, when most of the observations are in the less variable group (here the null group, as the variance increases with the mean), then the variance component is underestimated and

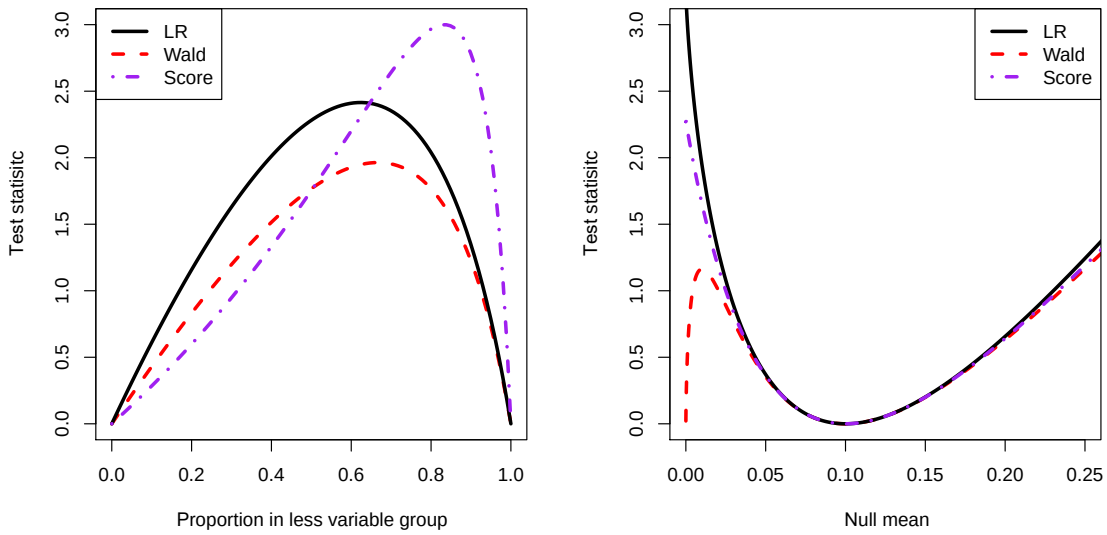
hence test statistics are inflated. This effect is far more pronounced for the score test statistic than the Wald and likelihood ratio. This kind of unbalanced design is often observed in ecological datasets, for example the bush regeneration data (Section 1.1.1), which has two control sites and eight treatment sites. This effect is more generally observed when the predictors are skewed.

In Figure 7.2b, we plot the three statistics against the null group mean $\bar{y}_0$, with the alternate group mean held at $\bar{y}_1 = 0.1$, overdispersion $\phi = 2$, and a sample size of $N = 50$, with $n_0 = n_1 = 25$. The test statistics will be zero when $\bar{y}_0 = \bar{y}_1 = 0.1$ and should, for a good test, increase monotonically as $\bar{y}_0$ moves away from $\bar{y}_1$ in either direction. That is, as the difference between the groups increases, so does the power of the test. In the case of the score and likelihood ratio statistic, this is indeed the case. However, the Wald statistic approaches zero as $\bar{y}_0$ approaches the boundary of zero. Again, such low means are prevalent in multivariate abundance data. For example, in the bush regeneration data (Section 1.1.1), the order *Blattodea* only appears in one of the eight regenerated sites.

7.1.3 Simulation results

In Chapter 5 we specified a copula likelihood, which can be used to conduct likelihood ratio tests for multivariate discrete data, even when sample size is not large relative to the number of variables. In this section we compare this likelihood ratio test with the Wald and score test conducted with a GEE (Section 2.2), on data with properties (mean, overdispersion and correlation) derived from the bush regeneration data (Section 1.1.1). The data are simulated from a Gaussian copula with negative binomial marginal distributions.

To assess the effect of unbalanced design, we simulated from a bivariate model, with varying sample sizes in null and treatment groups. For all simulations we have a single covariate, a treatment effect, which is either in the direction of the main eigenvector of the covariance matrix between species or orthogonal to it. For each simulated dataset, we conduct a test for treatment effect with GEEs (using both Wald and score statistics, implemented in the `mvabund` package in R) and using Gaussian copulas with a likelihood ratio test. We estimate the likelihood for the



(a) **(b)**

Figure 7.2:

(a) Observed test statistic plotted against the proportion of observations in the less variable group for negative binomial with a treatment effect and $\bar{y}_0 = 0.1$, $\bar{y}_1 = 1$, $\phi = 2$ and $N = 10$. If the variance of the two groups was not relevant to power, we would expect to see this graph be symmetric around 0.5. When most of the observations are in the less variable group, the variance is underestimated and hence test statistics are inflated. This effect is far more pronounced for the score statistic than the Wald and likelihood ratio statistics.

(b): Test statistics plotted against the null group mean $\bar{y}_0$, with the alternate group mean held at $\bar{y}_1 = 0.1$, $\phi = 2$, $n_0 = n_1 = 25$. The test statistic will be 0 when $\bar{y}_0 = \bar{y}_1 = 0.1$ and should, ideally, increase monotonically as $\bar{y}_0$ moves away from $\bar{y}_1$ in either direction. In the case of the score and likelihood ratio statistic, this is the case, however the Wald statistic approaches 0 as $\bar{y}_0$ approaches the boundary of 0.

Gaussian copula model (Equation 5.1) with Algorithm 1, with negative binomial marginal distributions. We use the `mvabund` package to estimate the marginals, with the overdispersion parameter estimated by maximum likelihood (for further simulation detail see Appendix B.2.1). Note that all three tests estimate marginal models in the same way, the only difference is how covariance is estimated and used to conduct hypothesis testing for parameters.

The results of these simulations are displayed in Figure 7.3. The top two plots have more observations in the treatment (more variable) group, and we can see the score statistic has very low power in this situation, followed by the Wald test, with the likelihood ratio test faring the best. In the bottom two plots, more observations are in the control (less variable) group, and the score test outperforms the others. If the treatment effect is orthogonal to the main eigenvector of the covariance matrix (Figure 7.3, left), tests that estimate correlation (solid lines) outperform tests that assume independence (dotted lines), while the opposite is true when the treatment effect is along the main eigenvector (Figure 7.3, right).

These results are consistent with our expectations, as discussed in Section 7.1.2. The score and Wald test perform worse when more observations are in the less variable group, while the likelihood ratio test performs more consistently across unbalanced designs. In addition, estimating covariance improves power of all tests, except when the treatment effect is in the direction of the dominant eigenvector.

7.2 Model selection

7.2.1 Model selection for marginal models

Model selection is widely used in ecology (Burnham et al., 2011; Grueber et al., 2011). In the case of multivariate abundance data, many environmental variables are routinely collected, and interest is in which of these are related to the community of species or which would be useful in predicting communities at unsampled sites.

Model selection for marginal models for multivariate abundance data can be carried out using the GEE framework. However, this is complicated by the lack of an explicit

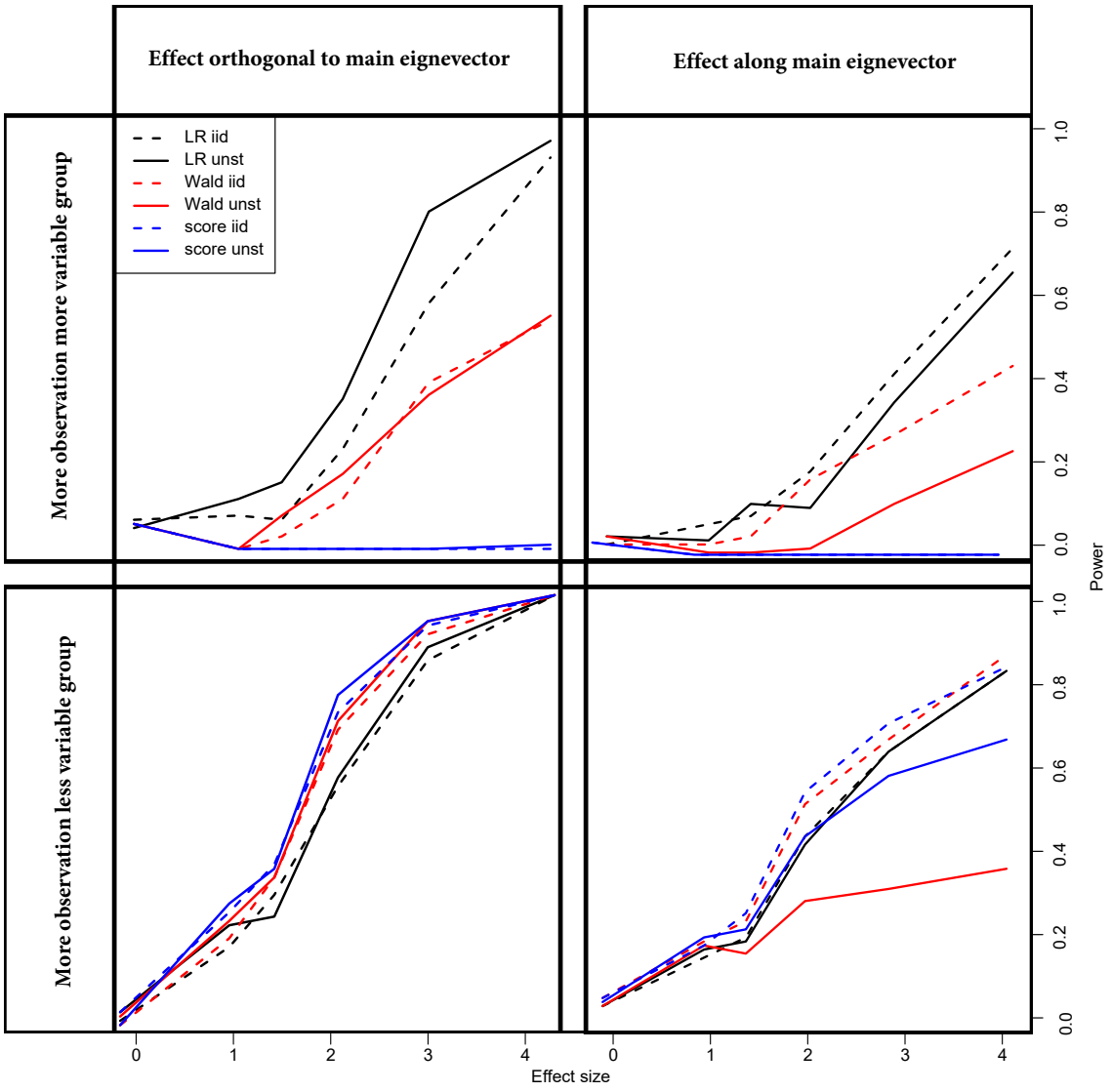


Figure 7.3: Results of power simulation for hypothesis tests when varying the direction of the treatment effect and the number of observations in less variable group. When the treatment effect is along the main eigenvector of the covariance matrix (right), tests assuming independence (- - -) are more powerful. Conversely, the opposite is true then the effect is orthogonal (left). When more of the observations are in the more variable group (top), the score test has very little power, while when more of the observations are in the less variable group (bottom), the score test outperforms the others.

likelihood unless independence is assumed. Model selection using likelihood based information criteria like Akaike information criterion (AIC; Akaike, 1974) and the Bayesian information criterion (BIC; Schwarz et al., 1978) can only be conducted under the assumption of independence, that is, adding univariate information criteria across species (AIC indep; Lyons et al., 2016). Another option is to implement criteria which use pseudo likelihoods (Pan, 2001a; Cantoni et al., 2005; Wang & Qu, 2009; Wang et al., 2012a; Cho & Qu, 2013) that assume independence (though the correlation matrix appears in the penalty term). Similarly to hypothesis testing (Section 7.1.1), we would expect criteria that assume independence to perform well when the covariate effects align with the direction of the main eigenvectors of the covariance matrix, and poorly when orthogonal to them.

Another option is to use non likelihood based model selection criteria in the GEE framework. The score information criterion (SIC; Stoklosa et al., 2014) does not assume independence in the likelihood estimate. It is based on the score statistic, and we expect it to have similar properties to the score test (see Section 7.1.2). We will explore how these criteria compare to AIC based on the copula likelihood (AIC copula).

7.2.2 Simulation study

We will explore the behaviour of the SIC, QIC (Pan, 2001a) and AIC with a Gaussian copula model for multivariate abundance data. We begin by defining these criteria. The QIC is defined as

$$QIC = -2\sum_{i=1}^{N}\sum_{j=1}^{P}\log f_j(y_{ij},\theta_j) + 2\text{tr}(\Omega\hat{\Sigma}(\theta))$$

Here $\hat{\Sigma}(\theta)$ is a sandwich estimator of $\text{cov}(\hat{\theta})$ given in Equation (2.4), and Ω is the naive estimator of covariance, given in Equation (2.3). Notice the covariance of parameters only enters into the penalty term, and not the likelihood. The AIC is defined as

$$AIC = -2\log L(\theta) + 2q,$$

where q is the number of variables in the model. The likelihood for the Gaussian copula model is defined in Equation (5.1). The SIC is defined as

$$SIC = -u(\theta_0)^T \Sigma^{-1}(\theta) u(\theta_0) + 2q.$$

Unlike the other criteria, the SIC is not defined for one model, but between two models, a null model with $\theta = \theta_0$ and an alternate model where some components of θ which are zero under the null model are allowed to vary.

We simulated data to mimic the properties of the spider dataset (Section 1.1.2) with means, overdispersion and the number of species and sites derived from the data. Response was a log linear function of the predictor. To test for sensitivity to the direction of covariate effects, we simulated environmental effects along all the eigenvectors of the covariance matrix, and measured the proportion of simulations for which the model selection criteria chose the correct model. For all marginal models (for both GEEs and copula models) we fit negative binomial distributions with the `mvabund` package in R. Wald and score tests were also conducted using the `mvabund` package, while copula models were estimated with Algorithm 1. For further simulation details see Appendix B.2.2. Results for the AIC assuming independence of species (AIC indep), QIC, AIC using the copula likelihood (AIC copula) and SIC are shown in Figure 7.4. There are $P = 12$ eigenvectors, with eigenvalues of a range of magnitudes. As expected (Section 7.1.1) the AIC assuming independence, and the QIC, which assumes independence in the quasi likelihood term, has similar power regardless of the direction of the covariate effect. On the other hand, criteria that estimate dependence (SIC and AIC copula) are more powerful when the effect of treatment is along a less dominant eigenvector (those with smaller eigenvalues).

To investigate the differences between the copula AIC and the SIC for unbalanced and skewed predictors, we simulated from a bivariate negative binomial model with a predictor that acted to increase the abundance of both simulated species. In this scenario we expect the SIC will be more powerful for positively skewed predictors, or for unbalanced binary predictors when most observations are in the less variable group. In these circumstances the SIC underestimates the variance component (Section 7.1.2).

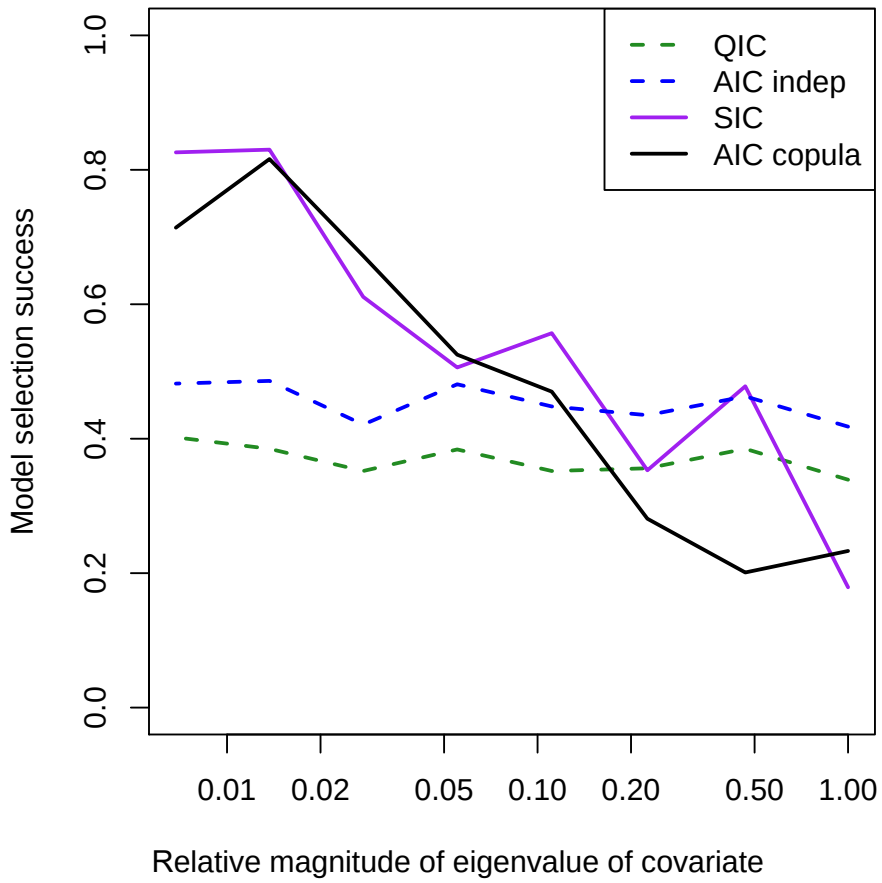


Figure 7.4: Model selection success (proportion of simulations that chose the correct model) plotted against the direction of covariate effect. Model selection procedures which assume independence (QIC and sum of AIC) do not react to the direction of the covariate effect, while the SIC and copula are more powerful when the covariate effect is along the eigenvectors with smaller eigenvalues (on the left), as this is the direction where data is least variable.

We again simulate data based on the properties of the hunting spider data, however the data we simulate are bivariate, so the first two species only are used. We generate either a binary predictor, with a set proportion (between 0 and 1) of observations in the less variable (null) group, or a skewed predictor (for a range of skewness values) with zero mean and unit variance. We then generate data from a negative binomial GLM with the relevant predictor. We carry out model selection using each criterion, with candidate models having either the correct predictor or no predictor. For more simulation details see Appendix B.2.3. Figure 7.5 shows the proportion of simulations which chose the correct model plotted against the proportion of observation in the less variable group (left) and the skewness of the predictor (right). The skewness of the predictor (left) has no effect on the likelihood based criteria (AIC copula), while the SIC is more powerful when the predictor is positively skewed (right). For the unbalanced sampling design (left), the copula AIC is more powerful for a balanced design, while the SIC is more powerful when most of the observations are in the less variable group. The main advantage of the SIC is computational, as it requires fitting only the least complex model at each stage of the forward selection path. However, unlike the other criteria discussed, it cannot carry out all subsets selection, because the criterion is defined in terms of the sequence in which models are added (Stoklosa et al., 2014).

7.2.3 Model selection for covariance models

We have introduced inference for multivariate abundance data with Gaussian copulas and covariance modelling. Covariance models often sit in a larger class of models, and model selection is typically necessary on this aspect of the model as well. For example, in factor analysis, model selection is needed to chose the number of factors, while for graphical models the sparsity of the graph is chosen using model selection. Using a Gaussian copula likelihood we can select for both marginal and covariance parameter models using traditional information criteria like AIC and BIC. This allows us to better investigate the patterns in covariance between species, as detailed in Chapter 4.

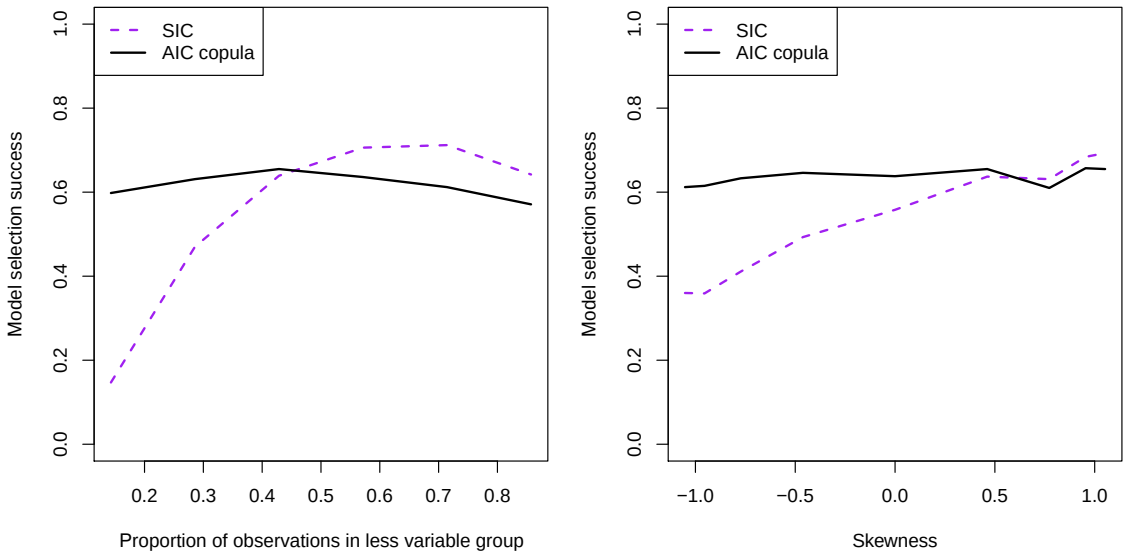


Figure 7.5: Model selection success (proportion of simulations which chose the correct model) plotted against the the proportion of observations in the less variable group for a binary covariate (left) and skewness of a continuous covariate (right). Likelihood based model selection procedures, such as the AIC using a Gaussian copula model (—), tend to do best for balanced samples (when the proportion in the less variable group is close to 0.5), and are not impacted by skewness (right). When most of the observations are in the less variable group (left), and when predictors are positively skewed (right), the SIC (- - -) is more powerful.

7.3 Data analysis

We conduct Wald, score and likelihood ratio tests on the bush regeneration dataset (Section 1.1.1). The resulting p-values (Table 7.1) for the Wald and likelihood ratio tests indicate evidence of an effect of treatment, while the score test does not. We expect this pattern when there is an unbalanced design with most of the observations in the more variable group (Figure 7.2b). As the regeneration data are counts modelled with a negative binomial distribution, the more variable group is the group with the higher mean, which for most orders is the regeneration group (Figure 1.3), with the notable exception of *Blattodea*.

Test	p-value
GEE Wald	0.028
GEE score	0.307
Copula LR	0.026

Table 7.1: Results of hypothesis tests using the Wald and score test with GEEs and the Gaussian copula likelihood ratio. All tests are conducted by re-sampling due to small sample sizes.

Next we conduct model selection on the spider dataset (Section 1.1.2), adding one variable at a time to an intercept model. The change in the information criteria which estimate correlation (AIC copula and SIC) is listed in Table 7.2. For the two most skewed variables, cover of bare sand and cover of herb layer, we observe very similar change in AIC relative to an intercept model, but quite different changes in SIC values. This is consistent with Figure 7.5, which shows that SIC is sensitive to the skewness of predictors, while AIC is not.

7.4 Discussion

We have shown that Gaussian copula models allow us to conduct likelihood based inference, both for hypothesis testing and model selection, on discrete multivariate abundance data. Currently used inference based on approximations to the likelihood, like Wald and score tests, and related information criteria, have undesirable power properties for such data, and the Gaussian copula likelihood often outper-

Variable	Δ AIC	SIC	Skewness
dry soil mass	-76.1059	-34.5244	-0.67706
cover of bare sand	-41.5617	-8.54002	0.792439
cover of fallen leaves	-46.6372	9.964125	0.631565
cover of moss	-8.44453	-26.7662	0.206096
cover of herb layer	-44.6164	-36.4822	-0.75924
reflection	-42.5855	3.321501	-0.3075

Table 7.2: The change in AIC and SIC when adding each variable individually to an intercept model, and skewness of each variable. The two variables with the largest difference in skewness (bare sand and herb layer) have similar AIC values, but very different changes in SIC values. This is consistent with the SIC being sensitive to skewness.

forms these methods. In addition, we can conduct model selection on the covariance models to better understand the patterns in covariance between variables (species), which is not possible using GEE based approaches. Both hypothesis testing and model selection require only a small addition to existing GEE models to estimate covariance matrices, and can be done with minimal additional computational overhead.

Variable selection methods not discussed here can also be implemented with a Gaussian copula likelihood. For very high dimensional problems, sure independence screening procedures (Fan & Lv, 2008) use measures of dependence between the response and each covariate separately to chose a model. For generalised linear models, the likelihood for a model fitting each covariate separately can be used to conduct such screening procedures (Fan & Song, 2010). A natural multivariate extension of this would be a copula likelihood, as described in this chapter.

Another well known variable selection method is the lasso (Tibshirani, 1996; Hastie et al., 2015). This applies a L_1 penalty to the coefficients in a regression, which encourages some of them to shrink to zero, thereby excluding them from the model. This has been applied to modelling multiple spices in the mixture model framework (Hui et al., 2015b), as well as for presence only data (point event data of species locations, Phillips et al., 2006; Renner & Warton, 2013). In this thesis we use

penalised likelihood to estimate sparse graphical Gaussian copula models (Section 5.4.1). A penalty could equally be applied to coefficients in the marginal models, to carry out model selection of covariates.

Chapter 8

Discussion

Copulas are a powerful, flexible method for marginal modelling of multivariate data. While many fields have made extensive use of copulas (*e.g.* finance and engineering; Cherubini et al., 2004; Genest & Favre, 2007), they remain largely unexplored in ecology (with some exceptions; Eskelson et al., 2011; de Valpine et al., 2014). This may be due in part to fast and accurate estimation methods for copulas with discrete margins only recently becoming available (Genz & Bretz, 2002; Masarotto & Varin, 2012). Given the need for flexible multivariate modelling in ecology, there is a lot of scope for copula models to be further implemented.

In this thesis we have demonstrated how copulas can be used to model sparse data and we have developed tools to study correlations and interactions between species, by adapting parsimonious models for covariance to discrete data (Chapter 5). Our focus in that chapter was on developing a flexible method, where existing covariance modelling algorithms could be used off the shelf with discrete data. The algorithm we present produces consistent estimates (Section A.2), however it does not jointly maximise all model parameters. An algorithm which is both flexible and which jointly maximises model parameters could be further explored. We expect such an algorithm to be more computationally intensive, though feasible. It would improve efficiency of estimates, but we expect such improvements may be small (Joe, 2005).

We presented, in Chapter 6, a novel way to visualise multivariate abundance data, which provides information about species interactions. The conditional dependence

relationships displayed by our method cannot be visualised with the dominant visualisation methods like nMDS, which use algorithmic approaches. This is an example where models can be used to answer more complex questions.

We have discussed the limitations of this method in the context of sparse data, which limit it to being an exploratory tool for multivariate abundance data. In cases when sample sizes greatly exceed the dimension, these models can be expected to find conditional dependence relationships with some confidence (Liu et al., 2012; Strobl et al., 2012). In future we would like to explore how measures of uncertainty in these relationships can be calculated, and how these are affected by dimension, sample size, and discreteness of the data.

Chapter 7 introduced inference methods for multivariate abundance data using Gaussian copulas with discrete margins, and demonstrated superior power properties relative to alternative methods, by using likelihood based hypothesis testing and model selection, and estimating correlations. While model selection can be carried out efficiently, due to small sample sizes, the likelihood ratio test we present, like the alternative methods, relies on residual re-sampling for inference. This makes the method (and others) quite computationally intensive. It would be of interest to explore alternative and less computationally intensive strategies.

8.1 Further extensions

Ecologists are very interested in how and why species and sites are related. Possible reasons for correlations already discussed include shared response to environmental covariates (both measured and unmeasured), temporal and spatial patterns, species traits, phylogeny, and interactions with common species (Warton et al., 2015a). We have not explored these all in detail in this thesis, however copula models could be adapted to model many of these mechanisms.

Spatial factor analysis has been proposed as a way to study multivariate abundance data with spatial latent variables (Wang & Wall, 2003; Thorson et al., 2015, 2016). These can model unobserved environmental covariates that are spatially smooth. Different latent factors can model spatial correlation at several scales (Ovaskainen

et al., 2016). An advantage of these, relative to standard factor analytic models, is that they improve prediction at unobserved locations, in a similar way to kriging (Stein, 2012). The method proposed in Chapter 5 can be straightforwardly extended to model spatial factors and make such prediction simple and fast.

For temporal data, vector autoregressive state space models can model species correlations in time. For sparse data they can be implemented as a dynamic factor analysis (Zuur et al., 2003), which reduces the number of parameters to be estimated in a similar way to spatial factor analysis, with factors that are smooth in time. Some current implementations assume Gaussian errors (e.g, Holmes et al., 2012), though hierarchical models that allow exponential response distributions are also available (Helske, 2014). Multivariate time series models can be built in the copula framework (Heinen & Rengifo, 2007; Brechmann & Czado, 2015), and these could be extended to model sparse data with a dynamic factor analysis structure, to better model sparse discrete multivariate data.

For heterogeneous data, mixtures of factor analysers (Zoubin et al., 1996) can simultaneously perform clustering and ordination of multivariate data. These models have been extended to non Gaussian responses with hierarchical Bayesian models (Hui, 2017). Mixtures of factor analysers are covariance models, and so the method in Chapter 5 can be extended to perform ordination and clustering with a marginal, likelihood based model.

Many datasets in ecology are collected in a structured way, for example several samples taken within sites, to estimate within site variation. This process induces correlation in the data, with samples within one site being more similar than samples from different sites. Most often this is modelled (for univariate responses) with random effects in hierarchical mixed models (McCulloch & Neuhaus, 2006). Multivariate models with parsimonious covariance structure and random effects have been proposed and applied to ecological data (Ovaskainen & Soininen, 2011), but methods to apply these to a wide range of data types (counts, biomass and ordinal) are not widely available. In contrast, copula models for discrete data could be straightforwardly extended to incorporate these sources of correlation, while still modelling between species correlations in a parsimonious way with, for example, la-

tent factors. For copula models, changing marginal distributions is trivial. Inference methods explored in Chapter 7 can be extended to carry out hypothesis testing and model selection in this context.

Another possible source of correlation we have not discussed in detail are phylogenetic similarities (Webb et al., 2002). Species that have similar phylogeny may be expected to inhabit more similar environments than those that are phenologically different. It is of interest to account for such correlations in modelling, as it can reduce the complexity of the species covariance matrix. This could be carried out by using phylogenetic distance as a basis for correlation functions, similar to the way spatial correlation is often modelled. Another option is to carry out covariance regression (Hoff & Niu, 2012). These models could provide additional information about the extent to which phylogeny induces correlations. Phylogeny has a complex relationship with species traits (Kraft et al., 2007; Best et al., 2013), and the interplay of these could be studied with copula models.

The focus of this thesis has been copula modelling of multivariate abundance data, which has, among other things, allowed us to apply covariance models not generally used in ecology to investigate species interactions. In particular in Chapter 6 we introduce a method to visualise species interactions using graphical models in a Gaussian copula framework. Graphical models could also be incorporated into a hierarchical framework more commonly used in ecology. Model of this type have been used in other areas, including RNA and microbial sequencing (Gallopin et al., 2013; Dangl & Jojic, 2015). Using marginal Poisson distributions, such models would need to be extended to handle the overdispersion generally found in multivariate abundance data.

This thesis has made significant advances to inferential tools for multivariate data in ecology. Copula modelling, combined with covariance models, can yield insights into patterns in covariance between species, and allow for parsimonious likelihood based inference. We provide, in this thesis, a basis for greater application of copula models in community ecology, as there is vast scope for this approach to yield further insights.

Appendix A

APPENDIX: PROOFS

A.1 Proof of Lemma 1

We aim to show the ith component of discrete Gaussian copula likelihood can be approximated by importance sampling with K sets of Dunn-Smyth residuals, that is

$$\begin{aligned} L_i(y_i; R_\theta) &= \int_{A_i} |R_\theta|^{-1/2} \exp\left(-\frac{1}{2} z_i^T (R_\theta^{-1} - I) z_i\right) du_i \\ &\approx \prod_{j=1}^{P} f_{ij}(y_{ij}) \sum_{k=1}^{K} c(z_i^k; R_\theta), \end{aligned}$$

where z_i^k is the kth sample of Dunn-Smyth residuals for observation i, $c(\cdot; R_\theta) = N_P(z_i^k; R_\theta)/\prod_{j=1}^{P} \phi(z_{ij}^k)$ is the Gaussian copula density with correlation matrix R_θ. The region of integration is given by $A_i = \cap_{j=1}^{P} \left[F_{ij}(y_{ij}^-|\beta_j, \psi_j), F_{ij}(y_{ij}|\beta_j, \psi_j)\right]$, $F_{ij}(y_{ij}^-|\cdot) = \lim_{x \to y_{ij}^-} F_{ij}(x|\cdot)$, $N_P(\cdot; R)$ is a multivariate Gaussian density with zero mean, unit variance and correlation matrix R, and $\phi(\cdot)$ is a univariate standard Gaussian density.

Here $f_{ij}(y_{ij}) = P(Y_{ij} = y_{ij})$ is the probability mass function if y_{ij} are discrete, and the probability density function $f_{ij}(y_{ij}) = \frac{\partial F_{ij}(y_{ij})}{\partial y_{ij}}$ when y_{ij} are continuous. We assume the $f_{ij}(y_{ij})$ are well defined, *i.e* identifiable, continuous and at least three times differentiable as function of θ.

We begin by noting the distribution of the randomised Dunn-Smyth residuals given

the data and marginal distributions is

$$g(z) = \frac{\prod_{j=1}^{P} \phi(z_{ij}^k)}{\prod_{j=1}^{P} f_{ij}(y_{ij})} \mathbb{1}_{B_i}, \tag{A.1}$$

where $B_i = \cap_{j=1}^{P} \left[\Phi^{-1}(F_{ij}(y_{ij}^-|\beta_j, \psi_j)), \Phi^{-1}(F_{ij}(y_{ij}|\beta_j, \psi_j))\right]$. We then approximate the ith likelihood component by first changing the variable of integration to z_i, then multiplying and dividing by the importance sampling distribution in A.1, and finally approximating this distribution with randomised Dunn-Smyth residuals.

$$\begin{aligned}
L_i(y_i; R_\theta) &= \int_{A_i} |R_\theta|^{-1/2} \exp\left(-\frac{1}{2} z_i^T (R_\theta^{-1} - I) z_i\right) du_i \\
&= \int_{B_i} N_P(z_i; R_\theta) dz_i \\
&= \int_{B_i} N_P(z_i; R_\theta) \frac{\prod_{j=1}^{P} f_{ij}(y_{ij})}{\prod_{j=1}^{P} \phi(z_{ij}^k)} g(z_i) dz_i \\
&= \prod_{j=1}^{P} f_{ij}(y_{ij}) \int_{B_i} \frac{N_P(z_i; R_\theta)}{\prod_{j=1}^{P} \phi(z_{ij}^k)} g(z_i) dz_i \\
&\approx \prod_{j=1}^{P} f_{ij}(y_{ij}) \sum_{k=1}^{K} \frac{N_P(z_i^k; R_\theta)}{\prod_{j=1}^{P} \phi(z_{ij}^k)} \\
&= \prod_{j=1}^{P} f_{ij}(y_{ij}) \sum_{k=1}^{K} c(z^k; R_\theta)
\end{aligned}$$

□

A.2 Consistency

We aim to prove the consistency of estimates obtained by estimating a Gaussian copula with discrete margins using Algorithm 1. We follow the standard proof of consistency for maximum likelihood found in (for example) Ferguson (1996). The standard proof proceeds by defining $\tau(\theta)$, which is maximised at the maximum likelihood estimate (MLE) $\hat{\theta}$,

$$\tau(\theta) = \log \frac{L_n(\theta)}{L_n(\theta_0)} = \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(y_i;\theta)}{f(y_i;\theta_0)},$$

where y_i is a P-vector of data at observation $i = 1, \ldots, N$. This quantity then converges to it's expectation under θ_0 by the strong law of large numbers,

$$\frac{1}{n} \sum_{i=1}^{n} \log \frac{f(y_i;\theta)}{f(y_i;\theta_0)} \xrightarrow{P} E_{\theta_0} \log \frac{f(y;\theta)}{f(y,\theta_0)}.$$

This expectation is equal to the negative of the Kullback - Leibler divergence,

$$E_{\theta_0} \log \frac{f(y;\theta)}{f(y,\theta_0)} = -K(\theta_0, \theta) < 0$$

unless $f(y,\theta) = f(y;\theta_0)$. Therefore the MLE maximises $\tau(\theta)$ (assuming identifiability), which converges to a function which is maximised by θ_0, from which $\hat{\theta} \xrightarrow{p} \theta_0$ follows. A difficulty in our case is that we are not using MLEs for estimation – Algorithm 1 is a two-step estimation procedure, where we estimate β from a marginal likelihood and then maximise the conditional likelihood given these parameter estimates. We wish to show that treating β as 'nuisance parameters', we can get consistent estimates of parameters of R in the covariance model.

Conditions

We assume mild regularity conditions, where conditions 1-6 are stated in Casella & Berger (2002) Chapter 10.

1. The observation $y_i \sim f(y, \beta, R)$ for $i = 1, \ldots, N$ are independent.
2. β is identifiable, *i.e.* if $\beta \neq \beta'$ then $f(y, \beta, R) \neq f(y, \beta', R)$.
3. The densities $f(y, \beta, R)$ have common support, and f is differentiable in β.
4. The parameter space Ω contains an open set ω of which the true parameter β_0 is an interior point.

5. For every y in $\mathcal{Y}$ the density $f(y, \beta, R)$ is continuous and at least three times differentiable in β, and $\int f(y, \beta, R)dy$ can be differentiated three times under the integral sign.

6. There exists an open subset of $\omega \in \Omega$ containing β_0 and an integratable function $M_r(y)$, such that for every $\beta \in \omega$ and $y \in \mathcal{Y}$
$$\left| \frac{\partial^3}{\partial^3 \beta_r} \log f(y, \beta, R) \right| \leq M_r(y)$$
for $r = 1, \ldots, \dim(\beta)$, where $E_{\beta_0}(M_r(y)) < \infty$

7. For $r = 1, 2, ..., \dim(\beta)$ there are bounded functions $V_r(y)$ such that in the neighbourhood of β_0 for any fixed R
$$\left(\frac{\partial}{\partial \beta_r} \log f(y_i, \beta, R) \right)^2 \leq V_r(y)$$
with $E_{\theta_0}(V_r(y)) < \infty$.

We proceed by defining the Gaussian copula likelihood for $\theta = (\beta, R)$ as
$$l_n(\theta) = \log L_n(\beta, R) = \frac{1}{n} \sum_{i=1}^{n} \log f(t_i; \beta, R),$$
where β is the $P \times K$ matrix, with $\beta_{j,k}$ being the coefficient for the kth covariate regressed on the jth variable. Let $\theta_0 = (\beta_0, R_0)$ be the true parameters, and $\hat{\beta}$ be the matrix of coefficients where the jth row is found by maximising the jth marginal likelihood, as in Algorithm 1 step 1;
$$\hat{\beta}_j = \text{argmax}_{\beta_j} \sum_{i=1}^{n} \log L_j(y_j, \beta_j). \tag{A.2}$$

Lemma 2. Equation A.2 is equivalent to using independence estimating equations in the GEE framework, which under conditions 1-6, are consistent (Liang & Zeger, 1986), so $\hat{\beta} \xrightarrow{P} \beta_0$.

Analogously to the standard maximum likelihood proof, the value $\hat{R}$ found by Algorithm 1 maximises $\tau'(R)$ where
$$\tau'(R) = \log \frac{L_n(\hat{\beta}, R)}{L_n(\hat{\beta}, R_0)} = \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(y_i; \hat{\beta}, R)}{f(y_i; \hat{\beta}, R_0)}.$$
We cannot use the law of large numbers directly to show this converges to its expectation under θ_0 as each summand of $\tau'(R)$ is a function of all the data, through $\hat{\beta}$. To prove this we first consider lemma 3.

Lemma 3.

$$\frac{1}{n} l_n(\hat{\beta}, R) \xrightarrow{P} E_{\theta_0} \log f(y, \beta_0, R)$$

Proof

Under conditions 1-7 it holds that for any fixed R the Taylor expansion of the standardised likelihood around β_0 is

$$\frac{1}{n} l_n(\hat{\beta}, R) = \frac{1}{n} l_n(\beta_0, R) + \frac{1}{n} (\hat{\beta} - \beta_0)^T \frac{\partial l_n(\beta, R)}{\partial \beta}\bigg|_{\tilde{\beta}}, \tag{A.3}$$

where $\tilde{\beta}$ is between $\hat{\beta}$ and β_0. By the Cauchy-Schwarz inequality, the last term is

$$\left|\left| \frac{1}{n} (\hat{\beta} - \beta_0)^T \frac{\partial l_n(\beta, R)}{\partial \beta}\bigg|_{\tilde{\beta}} \right|\right| \leq \frac{1}{n} ||\hat{\beta} - \beta_0|| \times \left|\left| \frac{\partial l_n(\beta, R)}{\partial \beta}\bigg|_{\tilde{\beta}} \right|\right|.$$

By Lemma 2, we know $||\hat{\beta} - \beta_0|| = o_p(1)$. We then look at the square of the last term

$$\left|\left| \left(\frac{\partial l_n(\beta, R)}{\partial \beta}\bigg|_{\tilde{\beta}} \right) \right|\right|^2 = \sum_{r=1}^{dim(\beta)} \left(\sum_{i=1}^{n} \frac{\partial}{\partial \beta_r} \log f(y_i, \beta, R)\bigg|_{\tilde{\beta}} \right)^2 = O_P(n^2),$$

which follows from the regularity conditions. Hence

$$\left|\left| \frac{\partial l(\beta, R)}{\partial \beta}\bigg|_{\tilde{\beta}} \right|\right| = O_P(n),$$

So the remainder term in equation A.3 is given by

$$\begin{aligned} \left|\left| \frac{1}{n} (\hat{\beta} - \beta_0)^T \frac{\partial l(\beta, R)}{\partial \beta}\bigg|_{\tilde{\beta}} \right|\right| &\leq \frac{1}{n} ||\hat{\beta} - \beta_0|| \times \left|\left| \frac{\partial l_n(\beta, R)}{\partial \beta}\bigg|_{\tilde{\beta}} \right|\right| \\ &= \frac{1}{n} o_P(1) O_P(n) = o_P(1). \end{aligned}$$

This in turn implies

$$\begin{aligned} \frac{1}{n} l_n(\hat{\beta}, R) &= \frac{1}{n} l_n(\beta_0, R) + \frac{1}{n} (\hat{\beta} - \beta_0)^T \frac{\partial l_n(\beta, R)}{\partial \beta}\bigg|_{\tilde{\beta}} \\ &= \frac{1}{n} l_n(\beta_0, R) + o_P(1). \end{aligned}$$

Hence for any R

$$\frac{1}{n} l_n(\hat{\beta}, R) \xrightarrow{P} E_{\theta_0} \log f(y, \beta_0, R)$$

□

Now we can return to the standard proof. We have

$$\tau'(R) = \log \frac{L_n(\hat{\beta}, R)}{L_n(\hat{\beta}, R_0)} = \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(y_i; \hat{\beta}, R)}{f(y_i; \hat{\beta}, R_0)} \xrightarrow{P} E_{\theta_0} \log \frac{f(y; \beta_0, R)}{f(y, \beta_0, R_0)}$$
$$= -K(\theta_0, \theta) < 0$$

unless $f(y, \theta) = f(y; \theta_0)$, and so $\hat{\theta} \xrightarrow{P} \theta_0$ and hence $\hat{R} \xrightarrow{P} R_0$.

A.3 Proof of equivalence of Algorithm 1 to EM algorithm

Here we show that Algorithm 1 is an example of a MCEM algorithm. As discussed in Definition 1, the EM algorithm iterates the E-Step and M-Step. In the E step we calculate the Q function

$$Q(\theta, \hat{\theta}^{(m)}) = \sum_{i=1}^{N} \int_{z_i} f(z_i|y_i; R_{\hat{\theta}^{(m)}}) \log f(z_i; R_\theta) dz_i.$$

which is the expectation of the log likelihood with respect to the conditional predictive distribution $f(z|y; R_{\hat{\theta}^{(m)}})$. Here z are the latent variables and y_i are the discrete data, both of dimension P, and $\hat{\theta}^{(m)}$ is the estimate of θ from the previous iteration. The M-sept maximises the Q function. The two steps are iterated until convergence.

E-Step: Calculate weights (Algorithm 1, step 4b)

For a Gaussian copula with discrete margins, $f(z_i; R_\theta) = N_P(z_i; R_\theta)$ and

$$\begin{aligned} f(y_i|z_i; R_\theta) &= f(y = y'|z = z'; R_\theta) \\ &= \begin{cases} 1 & \text{if } z_i \in B_i = \cap_{j=1}^{P} \left[\Phi^{-1}(F_{ij}(y_{ij}^-|\beta_j, \psi_j)), \Phi^{-1}(F_{ij}(y_{ij}|\beta_j, \psi_j))\right] \\ 0 & \text{otherwise.} \end{cases}. \end{aligned}$$

And so we obtain the conditional likelihood

$$\begin{aligned} f(z_i|y_i; R_\theta) &\propto f(y_i|z_i; R_\theta) f(z_i; R_\theta) \\ &= \mathbb{1}_{z_i \in B_i} N_P(z_i; R_\theta), \end{aligned}$$

which is the truncated multivariate normal distribution with covariance matrix R_θ.

To carry out an MCEM algorithm we need to sample from $f(z_i|y_i; R_{\hat{\theta}^{(m)}})$ at the mth iteration (Definition 1). We do this by first sampling Dunn-Smyth residuals, whose distribution is a truncated multivariate normal with identity covariance matrix (see equation A.1), and then weight observations accordingly. To obtain draws from $f(z_i|y_i; R_{\theta^{(m)}})$ by weighting samples from $g(z)$, the weights must be proportional to

$$w'_{ik}(R_{\theta^{(m)}}) = \frac{f(z_i^k|y_i; R_{\theta^{(m)}})}{g(z_i^k)} \propto \frac{N_P(z_i^k; R_{\theta^{(m)}})}{\prod_{j=1}^{P} \phi(z_{ij}^k)} = c(z_i^k; R_{\theta^{(m)}}).$$

In equation 5.4 we define

$$w_{ik}(R_{\theta^{(m)}}) = \frac{w'_{ik}(R_{\theta^{(m)}})}{\sum_{k=1}^{K} w'_{ik}(R_{\theta^{(m)}})} = \frac{c(z_i^k; R_{\theta^{(m)}})}{\sum_{k=1}^{K} c(z_i^k; R_{\theta^{(m)}})} \propto c(z_i^k; R_{\theta^{(m)}}),$$

so proportionality is maintained, and hence the distribution of the weighted sample. We therefore have weighted samples $(z_i^k, w_{ik}(R_{\theta^{(m)}}))$ distributed according to $f(z_i|y_i; R_{\theta^{(m)}})$. Now the function to be maximised can be written

$$\begin{aligned} Q(\theta, \hat{\theta}^{(m)}) &= \sum_{i=1}^{N} \int_{z_i} f(z_i|y_i; R_{\hat{\theta}^{(m)}}) \log f(z_i; R_\theta) dz_i \\ &\approx \sum_{i=1}^{N} \sum_{k=1}^{K} w_{ik}(R_{\hat{\theta}^{(m)}}) \log N_P(z_i^k; R_\theta) \\ &:= \tilde{Q}(\theta, \hat{\theta}^{(m)}) \end{aligned}$$

M-Step: Maximise $\tilde{Q}$ function (Algorithm 1, step 4a)

We now need to maximise the $\tilde{Q}$ function. Differentiating $\tilde{Q}$ with respect to θ we get

$$\frac{\partial Q(\theta, \hat{\theta}^{(m)})}{\partial \theta} \approx \sum_{i=1}^{N} \sum_{k=1}^{K} w_{ik}(R_{\hat{\theta}^{(m)}}) \frac{\partial \log N_P(z_i^k; R_\theta)}{\partial \theta}.$$

This is the same as equation 5.4, and is the function maximised by applying covariance modelling algorithms to weighted Dunn-Smyth residuals in step 4a.

In Algorithm 1 step 4 we iterate these two steps until convergence. □

Appendix B

Simulation detail

B.1 Chapter 5

B.1.1 Figure 5.1

- Convert hunting spider data to presence/absence.
- Calculate marginal means for each species (probability of presence).
- For each combination of sample size N and number of species P (see Figure 5.1), simulate 100 binary samples with estimated means and covariance matrix modelled with one latent variable (standard Gaussian). When samples dimension P differs from data, sample P means from hunting spider data with replacement.
- Use Algorithm 1 with $K = 200$ Dunn-Smyth residuals, binomial marginal distributions and factor analysis covariance structure estimated with the `factanal` function to estimate model; extract estimated covariance matrix.
- Use `lava.tobit` to estimate latent variables model and extract estimated covariance matrices.
- Fit marginal GLMs using the `glm` function with `family = binomial(link="probit")`. Extract Pearson residuals and generate one set of Dunn-Smyth residuals according to Definition 3 in Chapter 5.

- Use the `factanal` function to estimate one factor model with the Pearson and Dunn-Smyth residuals and extract estimated covariance matrices.
- Calculate Frobenius norm of all estimated covariance matrices and the true covariance matrix.
- Plot ratios of these norms.

B.1.2 Figure 5.2

- For hunting spider data, calculate marginal means for each species.
- Generate graph with the `huge.generator` function in `huge` package with 70% probability of conditional independence between each pair of species, extract covariance matrix.
- For each combination of sample size N and number of species P (see Figure 5.1), simulate 100 samples from the Poisson distribution with estimated marginal means and simulated covariance matrix. When samples dimension P differs from data, sample P means from with replacement.
- Use Algorithm 1 with $K = 200$ Dunn-Smyth residuals, Poisson marginal distributions and Graphical model estimated with the `huge` package, extract graph.
- Estimate graph structure with local Poisson model (Allen & Liu, 2013).
- Fit marginal GLMs using the `glm` function with `family=Poisson`. Extract Pearson residuals and generate one set of Dunn-Smyth residuals according to Definition 3 in Chapter 5.
- Use the `huge` package to estimate graph with the Pearson and Dunn-Smyth residuals.
- Shrinkage parameter for all models selected with the StARS criterion (Liu et al., 2010).
- For each estimated graph calculate proportion of correctly identified conditional dependence relationships (recovery rate).
- Plot ratios of recovery rates.

B.2 Simulation detail Chapter 7

B.2.1 Figure 7.3

- For bush regeneration data, fit marginal negative binomial models with `manyglm` function in `mvabund` package, and extract estimated mean and overdispersion for first two species.
- Treatment effect is positive and equal for both species, such that the control group has a smaller mean than the treatment group.
- Correlation between species is either 0.8 (such that treatment effect is along the main eigenvector of correlation matrix) or -0.8 (such that treatment effect is orthogonal to the main eigenvector of correlation matrix)
- With a total sample size (N) of ten, let binary predictor (treatment) have either two or eight observations in the treatment group, for a range of effect sizes.
- For each combination of correlation $(-0.8, 0.8)$, effect size and number of observations in null group $(2, 8)$ conduct 100 simulations.
- Carry out Wald and score tests for treatment using the `manyglm` function in `mvabund` package, estimating covariance with both independence assumption `cor.type="I"` and unstructured covariance `cor.type="R"`.
- Carry out likelihood ratio test, with likelihood estimated with Algorithm 1, $K = 200$ Dunn-Smyth residuals and unstructured covariance matrix.
- Carry out likelihood ratio test under independence assumption by adding log likelihood for each species.
- Calculate power for each test as the proportion of simulations for which the test rejects the null hypothesis.

B.2.2 Figure 7.4

- For hunting spider data, fit marginal negative binomial models with one covariate (presence of bare sand) using the `manyglm` function in `mvabund` package. Extract null mean, overdispersion and treatment effect size and direction.

- Set sample size (N) to 28 and number of species (P) to 12 as in hunting spider data.
- Create matrix of size P where one eigenvector is the treatment direction and the others are orthogonal to it and one another.
- Create vector of eigenvalues on log scale.
- For $k = 1, \cdots P$, assign the kth largest eigenvalue to the eigenvector in the direction of treatment effect and create correlation matrix from eigenvectors and eigenvalues.
- Simulate 1000 observations for each k with corresponding correlation matrix, negative binomial distribution using the `mvabund` package, and treatment effect size derived from data.
- Fit model with no covariate (null model) and with correct covariate (alternate model).
- Select best model using QIC, AIC indep, SIC and AIC copula.
- Calculate proportion of times each model selection criterion selected the correct model.

B.2.3 Figure 7.5

- For hunting spider data, fit marginal negative binomial models using the `manyglm` function in `mvabund` package. Extract null means and overdispersion for first two species.
- Set sample size (N) to 28, number of species (P) to 2, and correlation between species to 0.8.
- [**Figure 7.5a**] Create binary covariate with proportion of observation in less variable group at 9 values between 0 and 1.
- [**Figure 7.5b**] Create continuous covariate X with zero mean, unit variance and skewness at one of 9 values.
- Simulate 1000 observations for each scenario with negative binomial distribution.
- Fit model with no covariate (null model) and with correct covariate (alternate model).

- Select best model using SIC and AIC copula.
- Calculate proportion of times each model selection criterion selected the correct model.

Bibliography

Abegaz, F., & Wit, E. (2014). Penalized EM algorithm and copula skeptic graphical models for inferring networks for mixed variables. Tech. rep., University of Groningen.

Abegaz, F., & Wit, E. (2015). Copula Gaussian graphical models with penalized ascent Monte Carlo EM algorithm. *Statistica Neerlandica*, *69*(4), 419–441.

Agresti, A. (2010). *Analysis of ordinal categorical data*, vol. 656. John Wiley & Sons, New York.

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions*, *19*(6), 716–723.

Allen, E. A., Damaraju, E., Plis, S. M., Erhardt, E. B., Eichele, T., & Calhoun, V. D. (2012). Tracking whole-brain connectivity dynamics in the resting state. *Cerebral Cortex*, *24*(3), 663.

Allen, G., & Liu, Z. (2013). A local Poisson graphical model for inferring networks from sequencing data. *NanoBioscience, IEEE Transactions*, *12*(3), 189–198.

Allen, R. (1993). A permanent plot method for monitoring changes in indigenous forests: a field manual. *Manaaki Whenua Landcare Research*.

Anderson, T. W. (1962). *An introduction to multivariate statistical analysis*. John Wiley & Sons, New York.

Austin, M. (2002). Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, *157*(23), 101 – 118.

Baba, K., Shibata, R., & Sibuya, M. (2004). Partial correlation and conditional correlation as measures of conditional independence. *Australian and New Zealand Journal of Statistics*, *46*(4), 657–664.

Banerjee, O., Ghaoui, L. E., d'Aspremont, A., & Natsoulis, G. (2006). Convex optimization techniques for fitting sparse Gaussian graphical models. In *Proceedings of the 23rd international conference on Machine learning*, (pp. 89–96).

Best, R. J., Caulk, N. C., & Stachowicz, J. J. (2013). Trait vs. phylogenetic diversity as predictors of competition and community composition in herbivorous marine amphipods. *Ecology Letters*, *16*(1), 72–80.

Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern wisconsin. *Ecological Monographs*, *27*(4), 325–349.

Brechmann, E. C., & Czado, C. (2015). COPAR multivariate time series modeling using the copula autoregressive model. *Applied Stochastic Models in Business and Industry*, *31*(4), 495–514.

Brock, J. M., Perry, G. L., Lee, W. G., & Burns, B. R. (2016). Tree fern ecology in New Zealand: A model for southern temperate rainforests. *Forest Ecology and Management*, *375*, 112 – 126.

Brown, A. M., Warton, D. I., Andrew, N. R., Binns, M., Cassis, G., & Gibb, H. (2014). The fourth-corner solution – Using predictive models to understand how species traits interact with the environment. *Methods in Ecology and Evolution*, *5*(4), 344–352.

Brown, J. H., Whitham, T. G., Morgan Ernest, S. K., & Gehring, C. A. (2001). Complex species interactions and the dynamics of ecological systems: Long-term experiments. *Science*, *293*(5530), 643–650.

Brownsey, P. J., & Perrie, L. R. (2014). Flora of New Zealand - Ferns and Lycophytes. In *Breitwieser, I. Heenan, Wilton, A.D. Flora of New Zealand*. Manaaki Whenua Press, Lincoln.

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, *65*(1), 23–35.

Cantoni, E., Flemming, J. M., & Ronchetti, E. (2005). Variable selection for marginal longitudinal generalized linear models. *Biometrics*, *61*(2), 507–514.

Carrara, F., Giometto, A., Seymour, M., Rinaldo, A., & Altermatt, F. (2015). Inferring species interactions in ecological communities: a comparison of methods at different levels of complexity. *Methods in Ecology and Evolution*, *6*(8), 895–906.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., & West, M. (2008). High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, *103*(484), 1438–1456.

Casella, G., & Berger, R. L. (2002). *Statistical inference*, vol. 2. Duxbury Pacific Grove, California.

Cherubini, U., Luciano, E., & Vecchiato, W. (2004). *Copula methods in finance*. John Wiley & Sons, New York.

Cho, H., & Qu, A. (2013). Model selection for correlated data with diverging number of parameters. *Statistica Sinica*, *23*(2), 901–927.

Craig, P. (2008). A new reconstruction of multivariate normal orthant probabilities. *Journal of the Royal Statistical Society: Series B*, *70*(1), 227–243.

Dangl, J. L., & Jojic, V. (2015). Lear microbial interaction networks from metagenomic count data. In *Research in Computational Molecular Biology: 19th Annual International Conference, RECOMB 2015, Warsaw, Poland, April 12-15, 2015, Proceedings*, vol. 9029, (p. 32). Springer, New York.

Dauwels, J., Yu, H., Xu, S., & Wang, X. (2013). Copula Gaussian graphical model for discrete data. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, (pp. 6283–6287).

de Valpine, P., Scranton, K., Knape, J., Ram, K., & Mills, N. J. (2014). The importance of individual developmental variation in stage-structured population models. *Ecology Letters*, *17*(8), 1026–1038.

Deheuvels, P. (1979). La fonction de dépendance empirique et ses propriétés. un test non paramétrique dindépendance. *Acadmie Royale de. Belgique. Bulletin de la Classe des Sciences. 6e Srie*, *65*(6), 274–292.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, *39*(1), 1–38.

Dobek, A., Moliński, K., & Skotarczak, E. (2015). Power comparison of Raos score test, the Wald test and the likelihood ratio test in (2xc) contingency tables. *Biometrical Letters*, *52*(2), 95–104.

Dobra, A., & Lenkoski, A. (2011). Copula Gaussian graphical models and their application to modeling functional disability data. *The Annals of Applied Statistics*, *5*(2A), 969–993.

Dunn, P. K., & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, *5*(3), 236–244.

Dunstan, P. K., Foster, S. D., & Darnell, R. (2011). Model based grouping of species across environmental gradients. *Ecological Modelling*, *222*(4), 955–963.

Eskelson, B. N. I., Madsen, L., Hagar, J. C., & Temesgen, H. (2011). Estimating riparian understory vegetation cover with beta regression and copula models. *Forest Science*, *57*(3), 212–221.

Everitt, B. S. (1984). *An introduction to latent variable models*. Springer, New York.

Fan, J., Liu, H., Ning, Y., & Zou, H. (2016). High dimensional semiparametric latent graphical model for mixed data. *Journal of the Royal Statistical Society: Series B*, (In press).

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*, *70*(5), 849–911.

Fan, J., & Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, *38*(6), 3567–3604.

Favre, A., El Adlouni, S., Perreault, L., Thiémonge, N., & Bobée, B. (2004). Multivariate hydrological frequency analysis using copulas. *Water Resources Research*, *40*(1).

Ferguson, T. S. (1996). *A course in large sample theory*. Chapman & Hall, London.

Ferrier, S., & Guisan, A. (2006). Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology*, *43*(3), 393–404.

Foster, S., Givens, G., Dornan, G., Dunstan, P., & Darnell, R. (2013). Modelling biological regions from multi-species and environmental data. *Environmetrics*, *24*(7), 489–499.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441.

Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, *21*(11), 1129–1164.

Gallopin, M., Rau, A., & Jaffrzic, F. (2013). A hierarchical Poisson log-normal model for network inference from RNA sequencing data. *PLOS ONE*, *8*(10), 1–9.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, Cambridge.

Genest, C., & Favre, A. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, *12*(4), 347–368.

Genest, C., Ghoudi, K., & Rivest, L. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, *82*(3), 543–552.

Genest, C., & Neslehova, J. (2007). A primer on copulas for count data. *Astin Bulletin*, *37*(2), 475–515.

Genz, A., & Bretz, F. (2002). Comparison of methods for the computation of multivariate t probabilities. *Journal of Computational and Graphical Statistics*, *11*(4), 950–971.

Gotelli, N. J., & Ulrich, W. (2010). The empirical bayes approach as a tool to identify non-random species associations. *Oecologia*, *162*(2), 463–477.

Grueber, C., Nakagawa, S., Laws, R., & Jamieson, I. (2011). Multimodel inference in ecology and evolution: challenges and solutions. *Journal of Evolutionary Biology*, *24*(4), 699–711.

Gruhl, J., Erosheva, E. A., Crane, P. K., et al. (2013). A semiparametric approach to mixed outcome latent variable models: Estimating the association between cognition and regional brain volumes. *The Annals of Applied Statistics*, *7*(4), 2361–2383.

Guo, J., Levina, E., Michailidis, G., & Zhu, J. (2015). Graphical models for ordinal data. *Journal of Computational and Graphical Statistics*, *24*(1), 183–204.

Hambleton, R. K. (1991). *Fundamentals of item response theory*. Sage Publications, New York.

Harris, D. J. (2015). Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution*, *6*(4), 465–473.

Harris, D. J. (2016). Inferring species interactions from co-occurrence data with Markov networks. *Ecology*, *97*(12), 3308–3314.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC Press, Florida.

Hauck Jr, W. W., & Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, *72*(360a), 851–853.

Heinen, A., & Rengifo, E. (2007). Multivariate autoregressive modeling of time series count data using copulas. *Journal of Empirical Finance*, *14*(4), 564–583.

Helske, J. (2014). KFAS: Kalman filter and smoothers for exponential family state space models. *R package version*, *1*, 4–1.

Hilbe, J. M., Hardin, J., & Hardin, H. (2003). *Generalized estimating equations*. CRC Press, Florida.

Hoff, P. D., & Niu, X. (2012). A covariance regression model. *Statistica Sinica*, *22*(2), 729–753.

Höfling, H., & Tibshirani, R. (2009). Estimation of sparse binary pairwise markov networks using pseudo-likelihoods. *The Journal of Machine Learning Research*, *10*, 883–906.

Holmes, E. E., Ward, E. J., & Wills, K. (2012). MARSS: Multivariate autoregressive state-space models for analyzing time-series data. *The R Journal*, *4*(1), 11–19.

Holst, K. K. (2012). *lava.tobit: LVM with censored and binary outcomes*. R package version 0.4-7.

Holst, K. K., & Budtz-Jørgensen, E. (2013). Linear latent variable models: the lava-package. *Computational Statistics*, *28*(4), 1385–1452.

Huang, S., Li, J., Sun, L., Ye, J., Fleisher, A., Wu, T., Chen, K., & Reiman, E. (2010). Learning brain connectivity of Alzheimer's disease by sparse inverse covariance estimation. *NeuroImage*, *50*(3), 935 – 949.

Hui, F. K. C. (2016). boral–Bayesian ordination and regression analysis of multivariate abundance data in R. *Methods in Ecology and Evolution*, *7*(6), 744–750.

Hui, F. K. C. (2017). Model-based simultaneous clustering and ordination of multivariate abundance data in ecology. *Computational Statistics and Data Analysis*, *105*, 1 – 10.

Hui, F. K. C., Taskinen, S., Pledger, S., Foster, S. D., & Warton, D. I. (2015a). Model-based approaches to unconstrained ordination. *Methods in Ecology and Evolution*, *6*(4), 399–411.

Hui, F. K. C., Warton, D. I., & Foster, S. D. (2015b). Multi-species distribution modeling using penalized mixture of regressions. *The Annals of Applied Statistics*, *9*(2), 866–882.

Hui, F. K. C., Warton, D. I., Foster, S. D., & Dunstan, P. K. (2013). To mix or not to mix: comparing the predictive performance of mixture models vs. separate species distribution models. *Ecology*, *94*(9), 1913–1919.

Hui, F. K. C., Warton, D. I., Ormerod, J. T., Haapaniemi, V., & Taskinen, S. (2016). Variational approximations for generalized linear latent variable models. *Journal of Computational and Graphical Statistics*, (In press).

Ives, A. R., & Helmus, M. R. (2011). Generalized linear mixed models for phylogenetic analyses of community structure. *Ecological Monographs*, *81*(3), 511–525.

Jamil, T., & ter Braak, C. J. (2013). Generalized linear mixed models can detect unimodal species-environment relationships. *PeerJ*, *1*, e95.

Joe, H. (1995). Approximations to multivariate normal rectangle probabilities based on conditional expectations. *Journal of the American Statistical Association*, *90*(431), 957–964.

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, *94*(2), 401–419.

Jordano, P., Bascompte, J., & Olesen, J. M. (2003). Invariant properties in coevolutionary networks of plantanimal interactions. *Ecology Letters*, *6*(1), 69–81.

Kraft, N., Cornwell, W., Webb, C., & Ackerly, D. (2007). Trait evolution, community assembly, and the phylogenetic structure of ecological communities. *The American Naturalist*, *170*(2), 271–283.

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, *29*(2), 115–129.

Leathwick, J. R. (1995). Climatic relationships of some New Zealand forest tree species. *Journal of Vegetation Science*, *6*(2), 237–248.

Ledoit, O., & Wolf, M. (2003). Honey, i shrunk the sample covariance matrix. *UPF economics and business working paper*, (691).

Lee, J. D., & Hastie, T. J. (2015). Learning the structure of mixed graphical models. *Journal of Computational and Graphical Statistics*, *24*(1), 230–253.

Legendre, P., & Legendre, L. F. (2012). *Numerical ecology*. Elsevier, Amsterdam.

Lemonte, A. J., & Ferrari, S. L. (2012). Local power and size properties of the LR, Wald, score and gradient tests in dispersion models. *Statistical Methodology*, *9*(5), 537–554.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary mathematics*, *80*(1), 221–239.

Liu, H., Han, F., Yuan, M., Lafferty, J., & Wasserman, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, *40*(4), 2293–2326.

Liu, H., Lafferty, J., & Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *The Journal of Machine Learning Research*, *10*, 2295–2328.

Liu, H., Roeder, K., & Wasserman, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, & A. Culotta (Eds.) *Advances in Neural Information Processing Systems 23*, (pp. 1432–1440). Curran Associates, Inc.

Lusk, C. H., & Laughlin, D. C. (2016). Regeneration patterns, environmental filtering and tree species coexistence in a temperate forest. *New Phytologist*, (In press).

Lyons, M. B., Keith, D. A., Warton, D. I., Somerville, M., & Kingsford, R. T. (2016). Model-based assessment of ecological community classifications. *Journal of Vegetation Science*, *27*(4), 704–715.

Masarotto, G., & Varin, C. (2012). Gaussian copula marginal regression. *Electronic Journal of Statistics*, *6*, 1517–1549.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. CRC Press, Florida.

McCulloch, C. E., & Neuhaus, J. M. (2006). *Generalized Linear Mixed Models*. John Wiley & Sons, New York.

McLachlan, G. J., & Krishnan, T. (1997). *The EM algorithm and extensions*. John Wiley & Sons, New York.

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, *34*(3), 1436–1462.

Meng, Z., Eriksson, B., & Hero, A. (2014). Learning latent variable Gaussian graphical models. In T. Jebara, & E. P. Xing (Eds.) *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, (pp. 1269–1277). JMLR Workshop and Conference Proceedings.

Miwa, T., Hayter, A., & Kuriki, S. (2003). The evaluation of general non-centred orthant probabilities. *Journal of the Royal Statistical Society: Series B*, *65*(1), 223–234.

Mohammadi, A., & Wit, E. C. (2015). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Analysis*, *10*(1), 109–138.

Morales-Castilla, I., Matias, M. G., Gravel, D., & Arajo, M. B. (2015). Inferring biotic interactions from proxies. *Trends in Ecology and Evolution*, *30*(6), 347 – 356.

Morueta-Holme, N., Blonder, B., Sandel, B., McGill, B. J., Peet, R. K., Ott, J. E., Violle, C., Enquist, B. J., Jrgensen, P. M., & Svenning, J.-C. (2016). A network approach for inferring species associations from co-occurrence data. *Ecography*, *39*(12), 1139–1150.

Murray, J. S., Dunson, D. B., Carin, L., & Lucas, J. E. (2013). Bayesian Gaussian copula factor models for mixed data. *Journal of the American Statistical Association*, *108*(502), 656–665.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A*, *135*(3), 370–384.

Nelsen, R. B. (1999). *An introduction to copulas*. Springer, New York.

Nikoloulopoulos, A. K. (2013a). Copula-based models for multivariate discrete response data. In P. Jaworski, F. Durante, & W. K. Härdle (Eds.) *Copulae in Mathematical and Quantitative Finance: Proceedings of the Workshop Held in Cracow, 10-11 July 2012*, (pp. 231–249). Springer, Berlin Heidelberg.

Nikoloulopoulos, A. K. (2013b). On the estimation of normal copula discrete regression models using the continuous extension and simulated likelihood. *Journal of Statistical Planning and Inference*, *143*(11), 1923–1937.

Ovaskainen, O., Abrego, N., Halme, P., & Dunson, D. (2016). Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods in Ecology and Evolution*, *7*(5), 549–555.

Ovaskainen, O., & Soininen, J. (2011). Making more out of sparse data: hierarchical modeling of species communities. *Ecology*, *92*(2), 289–295.

Pan, W. (2001a). Akaike's information criterion in generalized estimating equations. *Biometrics*, *57*(1), 120–125.

Pan, W. (2001b). On the robust variance estimator in generalised estimating equations. *Biometrika*, *88*(3), 901–906.

Peres-Neto, P. R., Olden, J. D., & Jackson, D. A. (2001). Environmentally constrained null models: site suitability as occupancy criterion. *Oikos*, *93*(1), 110–120.

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*(34), 231 – 259.

Pitt, M., Chan, D., & Kohn, R. (2006). Efficient Bayesian inference for Gaussian copula regression models. *Biometrika*, *93*(3), 537–554.

Pledger, S., & Arnold, R. (2014). Multivariate methods using mixtures: Correspondence analysis, scaling and pattern-detection. *Computational Statistics and Data Analysis*, *71*, 241–261.

Pollock, L. J., Morris, W. K., & Vesk, P. A. (2012). The role of functional traits in species distributions revealed through a hierarchical model. *Ecography*, *35*(8), 716–725.

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., Vesk, P. A., & McCarthy, M. A. (2014). Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (JSDM). *Methods in Ecology and Evolution*, *5*(5), 397–406.

Pourahmadi, M. (2013). *High-dimensional covariance estimation: with high-dimensional data*. John Wiley & Sons.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rao, C. R. (2009). *Linear statistical inference and its applications*. John Wiley & Sons, New York.

Ravikumar, P., Wainwright, M. J., & Lafferty, J. D. (2010). High-dimensional ising model selection using L1 regularized logistic regression. *The Annals of Statistics*, *38*(3), 1287–1319.

Renner, I. W., & Warton, D. I. (2013). Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics*, *69*(1), 274–281.

Rothman, A. J., Bickel, P. J., Levina, E., & Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, *2*, 494–515.

Rotnitzky, A., & Jewell, N. P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, *77*(3), 485–497.

Salvadori, G., & De Michele, C. (2004). Frequency analysis via copulas: Theoretical aspects and applications to hydrological events. *Water Resources Research*, *40*(12), 1–17.

Sánchez, B. N., Budtz-Jørgensen, E., Ryan, L. M., & Hu, H. (2005). Structural equation models: a review with applications to environmental epidemiology. *Journal of the American Statistical Association*, *100*(472), 1443–1455.

Schultz, N. A., Werner, J., Willenbrock, H., Roslind, A., Giese, N., Horn, T., Wøjdemann, M., & Johansen, J. S. (2012). MicroRNA expression profiles associated with pancreatic adenocarcinoma and ampullary adenocarcinoma. *Modern Pathology*, *25*(12), 1609–1622.

Schwarz, G., et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Shih, J. H., & Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, *51*(4), 1384–1399.

Sklar, M. (1959). *Fonctions de répartition à n dimensions et leurs marges*. Université Paris.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. CRC Press, Florida.

Smith, M. S., & Khaled, M. A. (2012). Estimation of copula models with discrete margins via Bayesian data augmentation. *Journal of the American Statistical Association*, *107*(497), 290–303.

Song, P. X. K., Li, M., & Yuan, Y. (2009). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics*, *65*(1), 60–68.

Stein, M. L. (2012). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, New York.

Stoklosa, J., Gibb, H., & Warton, D. I. (2014). Fast forward selection for generalized estimating equations with a large number of predictor variables. *Biometrics*, *70*(1), 110–120.

Strobl, R., Grill, E., & Mansmann, U. (2012). Graphical modeling of binary data using the LASSO: a simulation study. *BMC Medical Research Methodology*, *12*(1), 16.

Strong Jr, D. R., Simberloff, D., Abele, L. G., & Thistle, A. B. (2014). *Ecological communities: conceptual issues and the evidence*. Princeton University Press,Princeton.

Sutradhar, B. C., & Bartlett, R. F. (1993). Monte Carlo comparison of Wald's, likelihood ratio and Rao's tests. *Journal of Statistical Computation and Simulation*, *46*(1-2), 23–33.

ter Braak, C. J. (1986). Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis. *Ecology*, *67*(5), 1167–1179.

Thorson, J. T., Ianelli, J. N., Larsen, E. A., Ries, L., Scheuerell, M. D., Szuwalski, C., & Zipkin, E. F. (2016). Joint dynamic species distribution models: a tool for community ordination and spatio-temporal monitoring. *Global Ecology and Biogeography*, *25*(9), 1144–1158.

Thorson, J. T., Scheuerell, M. D., Shelton, A. O., See, K. E., Skaug, H. J., & Kristensen, K. (2015). Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range. *Methods in Ecology and Evolution*, *6*(6), 627–637.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, *58*(1), 267–288.

Væth, M. (1985). On the use of Wald's test in exponential families. *International Statistical Review / Revue Internationale de Statistique*, *53*(2), 199–214.

van Der aart, P., & Smeenk-Enserink, N. (1974). Correlations between distributions

of hunting spiders (Lycosidae, Ctenidae) and environmental characteristics in a dune area. *Netherlands Journal of Zoology*, *25*(1), 1–45.

ver Hoef, J. M., & Boveng, P. L. (2007). Quasi-poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology*, *88*(11), 2766–2772.

Walker, S. C., & Jackson, D. A. (2011). Random-effects ordination: describing and predicting multivariate correlations and co-occurrences. *Ecological Monographs*, *81*(4), 635–663.

Wang, F., & Wall, M. M. (2003). Generalized common spatial factor model. *Biostatistics*, *4*(4), 569–582.

Wang, L., & Qu, A. (2009). Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach. *Journal of the Royal Statistical Society: Series B*, *71*(1), 177–190.

Wang, L., Zhou, J., & Qu, A. (2012a). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, *68*(2), 353–360.

Wang, Y., Naumann, U., Wright, S., & Warton, D. (2012b). *mvabund: statistical methods for analysing multivariate abundance data*. R package version 3.8.0.

Warton, D. I. (2008). Which Wald statistic? Choosing a parameterization of the Wald statistic to maximize power in k-sample generalized estimating equations. *Journal of Statistical Planning and Inference*, *138*(10), 3269–3282.

Warton, D. I. (2011). Regularized sandwich estimators for analysis of high-dimensional data using generalized estimating equations. *Biometrics*, *67*(1), 116–123.

Warton, D. I. (2017). Why you cannot transform your way out of trouble for small counts. *Biometrics*.

Warton, D. I., Blanchet, F. G., OHara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C., & Hui, F. K. C. (2015a). So many variables: Joint modeling in community ecology. *Trends in Ecology and Evolution*, *30*(12), 766 – 779.

Warton, D. I., Foster, S. D., Death, G., Stoklosa, J., & Dunstan, P. K. (2015b). Model-based thinking for community ecology. *Plant Ecology*, *216*(5), 669–682.

Warton, D. I., Shipley, B., & Hastie, T. (2015c). CATS regression–a model-based approach to studying trait-based community assembly. *Methods in Ecology and Evolution*, *6*(4), 389–398.

Warton, D. I., Wright, S. T., & Wang, Y. (2012). Distance-based multivariate analyses confound location and dispersion effects. *Methods in Ecology and Evolution*, *3*(1), 89–101.

Webb, C. O., Ackerly, D. D., McPeek, M. A., & Donoghue, M. J. (2002). Phylogenies and community ecology. *Annual Review of Ecology and Systematics*, *33*(1), 475–505.

Wei, G. C., & Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, *85*(411), 699–704.

Wells, K., & O'Hara, R. B. (2013). Species interactions: estimating per-individual interaction strength and covariates before simplifying data into per-species ecological networks. *Methods in Ecology and Evolution*, *4*(1), 1–8.

Wiser, S. K., Hurst, J. M., Wright, E. F., & Allen, R. B. (2011). New Zealand's forest and shrubland communities: a quantitative classification based on a nationally representative plot network. *Applied Vegetation Science*, *14*(4), 506–523.

Xu, J. J. (1996). *Statistical modelling and inference for multivariate and longitudinal discrete response data*. Ph.D. thesis, University of British Columbia.

Yee, T. W. (2004). A new technique for maximum-likelihood canonical Gaussian ordination. *Ecological Monographs*, *74*(4), 685–701.

Yee, T. W., & Wild, C. J. (1996). Vector generalized additive models. *Journal of the Royal Statistical Society. Series B*, *58*(3), 481–493.

Yee, T. W., et al. (2010). The VGAM package for categorical data analysis. *Journal of Statistical Software*, *32*(10), 1–34.

Yuan, M., & Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, *94*(1), 19–35.

Zeger, S. L., & Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, *42*(1), 121–130.

Zhao, T., Liu, H., Roeder, K., Lafferty, J., & Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in R. *The Journal of Machine Learning Research*, *13*(1), 1059–1062.

Zoubin, Hinton, G. E., et al. (1996). The EM algorithm for mixtures of factor analyzers. Tech. rep., University of Toronto.

Zuur, A. F., Fryer, R. J., Jolliffe, I. T., Dekker, R., & Beukema, J. J. (2003). Estimating common trends in multivariate time series using dynamic factor analysis. *Environmetrics*, *14*(7), 665–685.