Models to relate species traits to environment: a hierarchical statistical approach

Tahira Jamil

Thesis committee

Thesis supervisor

Prof. dr. Cajo J.F. ter Braak
Personal Chair at Biometris
Wageningen University

Other members

Prof. dr. Joop H.J. Schaminée, Wageningen University
Prof. dr. ir. Alfred Stein, University of Twente, Enschede
Prof. dr. Herbert J.A. Hoijtink, Utrecht University
Dr. Mark J. de Rooij, Leiden University

This research was conducted under the auspices of the Graduate School of "Production Ecology & Resource Conservation"

Models to relate species traits to environment: a hierarchical statistical approach

Tahira Jamil

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. dr. M.J. Kropff,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday 11 January 2012
at 11 a.m. in the Aula.

Tahira Jamil

Models to relate species traits to environment: a hierarchical statistical approach

Thesis, Wageningen University, Wageningen, NL (2012)
With references, with summaries in English and Dutch

ISBN 978-94-6173-139-5

To my parents

Contents

Chapter 1

General introduction

In recent years the use of species distribution models by community ecologists has increased considerably. A central issue of community ecology is to understand species–environment relationship (Guisan and Zimmermann 2000). Many theories on community assembly assume that the effects of environmental factors are important in controlling and shaping species composition (Weiher and Keddy 1995). For over a century, ecologists have attempted to determine the factors that control species distribution (Motzkin et al. 2002). The importance of environmental factors to explain species distribution was recognized in the early 19th century. Despite the increasing number of investigations on species distributions and their relationship to the environment during the past decade, our understanding of how environmental conditions shape species distribution is still far from complete.

Models that predict distributions of species by combining species data with environmental variables have much potential for application in conservation. In the last two decades, interest in species distribution modelling therefore has grown dramatically and a wide variety of modelling techniques have been developed (Guisan and Thuiller 2005). These models commonly utilize associations between environmental variables and species data to identify environmental conditions within which species can be conserved.

The quantification of species–environment relationships represents the core of modelling in community ecology. A central focus of community ecology is to understand and explain where and when particular species or groups of species occur and thrive, and where and when not. To be able to survive, species need to be adapted to the environment they live in. Species differ in what they require and can tolerate from the environment, due to differences in traits, and environmental conditions vary in space and time. The relationship between species traits and environmental factors is therefore crucial for understanding of species community assembly. Although the role of species traits in community assembly has received much recent interest (Weiher et al. 1998, Lavorel and Garnier 2002, Statzner et al. 2004, Cornwell and Ackerly 2009, Ozinga et al. 2004, Ozinga et al. 2005a, Shipley et al. 2006, He 2010, Ozinga et al. 2005b), little is known about the relationships between species traits and different environmental factors.

A trait is a well-defined property of organisms that is usually measured at the organism level and used comparatively across species, as the intra-species variation is usually much smaller than the

inter-species variation (McGill et al. 2006). Several empirical studies have shown that species traits are associated with habitat conditions (Townsend and Hildrew 1994, Townsend et al. 1997, Pöyry et al. 2008) with the implication that traits were developed in the evolutionary process to enable species to adapt to the habitat and landscape characteristics in which the species occur (Southwood 1977, Ackerly 2003). Table 1 presents a small selection of trait-environment studies in recent years and shows the diversity in aim, taxa and statistical method used.

The increasing availability of species trait data is an extra stimulant for the growing interest of ecologists in the analysis of species functional trait responses to environmental conditions (Weiher et al. 1999, Violle et al. 2007). Despite this increasing interest, our knowledge of species community assembly is still hampered by lack of sound statistical methods for quantifying the effect of species traits on community assembly (Dray and Legendre 2008). Most studies on trait-environment relationship, especially model-based ones, are limited to a single species or to the effect of species traits on the performance of individual species. The idea of using species traits to study the properties of ecological communities is not new but currently a major research focus in ecology (McGill et al. 2006). However, knowledge of the effect of the relationship between species traits and environmental factor on species community assembly is still limited. This is partially because the link between species traits and environmental factors is mostly conjectured and limited to correlate them (McGill et al. 2006, Vile et al. 2006). Yet, species are seldom affected by only one environmental variable but experience different environmental factors simultaneously and also species often interact with other species in many ways. One of the major tasks of ecological studies is to analyse the response of community composition to environmental conditions and this often requires the use of multivariate analyses (Dray et al. 2003). Therefore there is an immense need for statistical methods that can link the environmental factors to traits in such multispecies communities. Species traits are also presumed have much predictive value for where and when a particular species or group of species appears or disappears. The central theme of this thesis is to develop models for species distribution that integrate the trait-environment relationships.

Trait-Environment relationships

Typical data in community ecology are arranged in two data tables (Fig. 1): a table **Y** recording the occurrence and abundance of numerous species in sites and a table **X** recording habitat and other site characteristics, i.e. the values or states of numerous environmental variables at the sites. Principal component analysis and correspondence analysis are ordination methods for analyzing a single table. Co-inertia analysis (Dolédec and Chessel 1994), redundancy analysis (Rao 1964), canonical correspondence analysis (ter Braak 1986) and canonical correlation analysis (Hotelling 1936) are multivariate methods for coupling two tables. Data on traits of species adds a third table **Z** (Fig. 1). For analysis of the three tables different approaches are used. Some authors combined

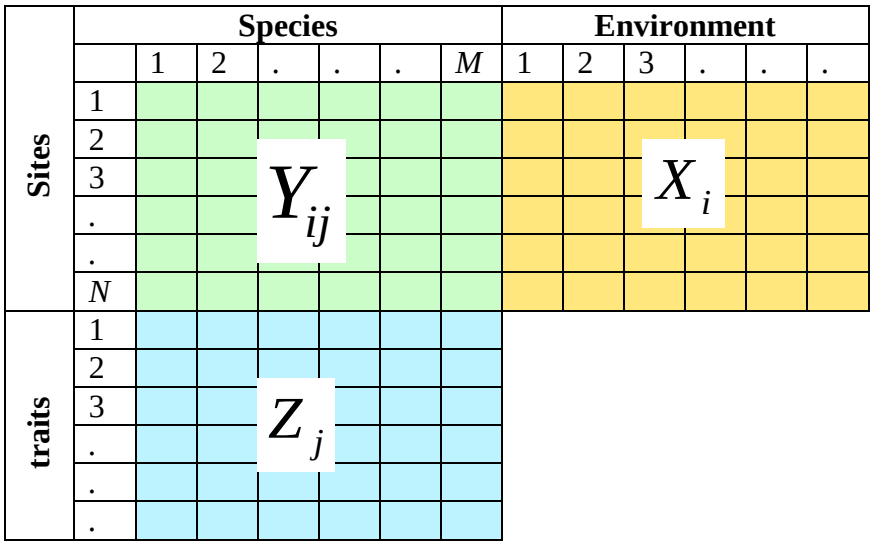

Fig. 1. Y is a sites × species table, X is a sites × environment table and Z is a species × trait table.

the sites × species table **Y** and species × trait table **Z** in to a sites × trait table which is then related to the sites × environment table **X** by standard statistical methods (Díaz et al. 1992, Sonnier et al. 2010). The combined table contains, for a quantitative trait, the mean across organisms at each site, for example the mean seed size, and for a qualitative trait, the numbers or percentages of organisms belonging to each category. This can also be done the other way round in which sites × species table **Y** and sites × environment table **X** are combined into a species × environment table which is then related to the species × trait table **Z** by standard (multivariate) statistical methods. The combined table may consist of mean environmental values per species or the percentage of individual of a species in each category of qualitative environmental variable. Legendre et al. (1997) and Dray and Legendre (2008) integrated these two steps in to one and called it the fourth-corner problem. The fourth corner is the lower-right gap in Fig. 1; it is the matrix to be constructed: the trait × environment matrix. For quantitative traits and environmental variables, the corner contains correlations between traits and environmental variables. The problem that a trait is measured on species and an environment variable on sites is circumvented by considering each number in each number in the sites × species table **Y** as a count of individuals of a particular species. Each individual has an attached trait value, namely that of the species to which the individual belongs, and an attached value of the environmental variable, namely the value of the site in which the individual occurs. Using individuals as cases, the Pearson correlation between the trait and the environmental variable is then calculated. The result can also be interpreted as a weighted Pearson correlation with the weight being the count. This interpretation is useful to generalize the method to sites × species tables with non-integer elements, to show that the fourth corner method ignores the zero values in the sites × species table **Y** and to show that negative

values are not allowed. Dray and Legendre (2008) examined six permutation-based methods to test the statistical significance of the trait-environment relationship, but none of them truly controlled the type I error. The fourth-corner method is descriptive; it is not a real modelling technique. The multivariate version of the fourth-corner problem is RLQ ordination (Dolédec et al. 1996) which has been used for selecting the best traits in species functional trait analyses (Bernhardt-Römermann et al. 2008).

The linear trait-environment (LTE) method (Cormont et al. 2011) was developed as a counterpart to the fourth corner method to enable usage of negative values in the sites × species table **Y**. In LTE, the trait-environment correlation is defined as the Pearson correlation between the species-specific regression coefficient and the species trait. The LTE method is a least squares method and the significance of the relationship is tested by a permutation test with a permutation strategy that does control the type I error. This strategy is a slight adaptation of one the permutation methods proposed for the fourth corner problem (Cormont et al. 2011).

Shipley et al. (2006) used the maximum entropy principle (MaxEnt) in an innovative way to predict microscopic features of communities (species abundance in sites) from macroscopic ones (their profiles of traits of species in the sites × trait table, where site refers to community). Interestingly, the result is a logistic model relating abundances to traits, as used in logistic regression but fitted in a different way (He 2010) and without environmental variables. The logistic regression model, which is the workhorse for statistical analysis of presence-absence data, can thus (also) be motivated from the maximum entropy principle. Ozinga et al. (2005a) started from the multiple logistic regression method and used it to quantify the effect of functional traits in a way that accounts for spatial variation in the composition of the local species pool. Their method assumes that species records within sites are independent (Hosmer and Lemeshow 2000), thus commits pseudo-replication (Hurlbert 1984, Crawley 2002). In applying generalized linear models, such as logistic regression, researchers often ignore the hierarchical structure of the data thereby producing incorrect variance estimates and increasing the likelihood of committing type I error (Wagner and Fortin 2005, Gillies et al. 2006).

Research Objectives:

The central focus of the thesis is how to quantify the relation of species traits with the environment via data on species occurrence and abundance in sites, species traits and the environmental characteristics. The research objectives when modeling species, traits and environments were:

- to develop a statistically sound and extendable framework for examining trait-environment relationships, and in particular,

- to develop a model-based approach modelling species data in relation to trait and environment data, taking into account the facts that species data are often in presence-absence form and that species often respond nonlinearly, even unimodal, to environmental change
- to develop methods to identify which traits and environmental factors best explain the distribution of species in space (and time)
- to evaluate how effective the newly developed methods are compared with existing methods of analysis.

Thesis outline

The thesis develops the methods to address the above questions by developing Bayesian and hierarchical models that utilize the trait information efficiently and that are able to automatically select the relevant traits and characteristics.

1. **Chapter 2** introduces a Generalized linear mixed model (GLMM) approach to identify which species traits and environmental variables best explain the species distribution in space and time, and which traits are significantly correlated with environmental variables. The GLMM approach is illustrated on a presence-absence version of the Dune Meadow data.

2. Niche theory predicts that species occurrence and abundance shows non-linear, unimodal relationships with respect to environmental gradients (Austin et al. 1984). The simplest symmetric unimodal species response to model unimodal relationships is the Gaussian response curve with three interpretable parameters (optimum, tolerance and height) that characterize the ecological niche of a species. Unimodal models, such as the Gaussian (logistic) model, are however much more difficult to fit to data than linear ones, particularly when also species traits are to be taken into account. **Chapter 3** develops a Bayesian approach to model unimodal species response to environment with submodels that relate the three niche parameters to species traits. The approach is illustrated with an application to phytoplankton communities.

3. Many studies fail to test for unimodal response. Thus straight-line relationships are often fitted without justification. **Chapter 4** studies the suitability of the GLMM approach for detecting unimodality of species response along an environmental gradient and suggests a graphical tool and a statistical test for testing unimodality. The efficacy of GLMM to analyze unimodal data when the niche widths are not very different among species is illustrated by comparing the GLMM approach with an explicit unimodal model approach on simulated data and real data that show unimodality.

4. **Chapter 5** studies the selection properties of Type II maximum likelihood (empirical bayes) in linear models with individual variance components for predictors. In a Bayesian framework, the variance components are estimated by using empirical bayes or, equivalently, by maximizing the marginal likelihood (type-II maximum likelihood) (Berger 1985).

5. **Chapter 6** develops the model selection method that is used Chapter 2 in more detail in a linear mixed model context. The method is called tiered forward selection. Using data from a mesocosm experiment, the linear mixed model with the tiered forward selection is compared with Type-II ML and existing methods for detecting trait-environment relationships that are not based on mixed models, namely the fourth corner method and the linear trait-environment method (LTE).

6. **Chapter 7** provides a summarizing discussion of the methods and their applicability in other research areas. It also gives limitations of methods and future possible research direction in species distribution modeling.

Table 1. Selected ecological studies on trait-environment relationships

Aim/topic	Taxa	Method	Reference
Analyze the effects of species trait variables on the accuracy of bioclimatic envelope models	Butterfly atlas data	Generalized additive models (GAMs)	(Pöyry et al. 2008)
Relating coral species traits to environmental conditions in the Jakarta Bay/Pulau Seribu reef system, Indonesia	Coral species	RLQ	(Rachello-Dolmen and Cleary 2007)
To quantify the effect of functional traits to accounts for spatial variation in the composition of the local species pool	Vascular plant	Multiple logistic regression	(Ozinga et al. 2005a)
Using functional traits to assess the role of hedgerow corridors as environmental filters for forest herbs	Herbs	Fourth-corner	(Poff et al. 2006)
Relating species traits to environmental variables in Indonesian coral reef sponge assemblages	Sponge species	RLQ	(de Voogd and Cleary 2007)
Morphology and life traits of species are related to the main underlying axes of environmental variability of their habitats	Beetle	RLQ	(Ribera et al. 2001)
Test the associations between biological and phyto geographical characteristics of species	Vascular flora	Fourth-corner	(Charest et al. 2000)
Selecting indicator traits for monitoring land use impacts	Birds	RLQ	(Hausner et al. 2003)
Analyze which characteristics of the species were significantly related to time since fire and distance from the forest at the sites	Forest tree	Fourth-corner	(Hooper et al. 2004)
The relationships between biological traits of macroinvertebrates and environmental characteristics with contrasting physical, chemical or landscape level attributes	Macroinvertebrates	RLQ	(Díaz et al. 2008)
To test for differences in the distribution of phylogenetic groups between the biofilm base and streamers	Bacteria	Fourth-corner revisited	(Besemer et al. 2009)
Using life-history traits to explain bird population responses to changing weather variability	Birds	LTE	(Cormont et al. 2011)

Chapter 2

Selecting traits that explain species-environment relationships: a Generalized Linear Mixed Model approach

Tahira Jamil, Wim A. Ozinga and Cajo J. F. ter Braak

(Submitted to Journal of Vegetation Scienc)

Abstract

Quantification of the effect of species traits on the assembly of communities is challenging from a statistical point of view. A key question is how species occurrence and abundance can be explained by the traits values of the species and the environmental values at the sites.

Using a sites × species abundance table, a site × environment data table and a species × trait data table, we address this question by a novel Generalized linear mixed model (GLMM) approach. We use numerical simulation to evaluate the testing procedure.

The GLMM can be used to identify which species traits and environmental variables best explain the species distribution in space and time, and which traits are significantly correlated with environmental variables.

We illustrate the approach on a presence-absence version of the Dune Meadow data and find that the species presence is best explained by moisture and manure of the meadows in combination with the Ellenberg's species traits coding for moisture, nitrogen and light requirements.

The GLMM overcomes the problem of pseudo-replication and heteroscedastic variance by including sites and species as random factors. The method is equally well applicable to presence-absence data as to count and multinomial data.

Key-words: community assembly; environmental gradient; trait-environment relationship; functional ecology; generalized linear mixed model; species traits

A central focus of community ecology is to understand and explain where and when particular species or groups of species occur and thrive, and where and when not. Species differ in what they require from the environment and environmental conditions vary in space and time. Differences in traits of species and differences in the environment must thus be part of the explanation. The role of species traits in community assembly has received much recent interest (Weiher et al. 1998, Lavorel and Garnier 2002, Statzner et al. 2004, Shipley et al. 2006, Cornwell and Ackerly 2009, He 2010). Quantification of the effect of traits on the assembly of communities turns out to be challenging from a statistical point of view (Dray and Legendre 2008).

Typical data in community ecology are arranged in two data tables: a table **Y** recording the occurrence and abundance of numerous species in sites and a table **X** recording habitat and other site characteristics, *i.e.* the values or states of numerous environmental variables at the sites (Fig.1). Such data are commonly used to study the relationships between species and environmental conditions, such as in species distribution models (Guisan and Zimmermann 2000, Guisan and Thuiller 2005) and direct and indirect gradient Analysis (ter Braak and Prentice 2004). Such models are powerful tools in investigating the possible consequences of changes in land-use and climate change on the distribution of species (Guisan and Zimmermann 2000, Raxworthy et al. 2003, Thuiller et al. 2005). They are also an important ingredient of conservation planning and management (Carroll et al. 2001, Raxworthy et al. 2003, Johnson et al. 2004).

These studies and models do not give much insight in why the species are distributed the way they are and why the species respond to changes in the way they do. Such insight might be gained by adding a third table **Z** (Fig. 1), a matrix with values and states of numerous species traits (Legendre et al. 1997, Dray and Legendre 2008). A trait is a well-defined property of organisms that is usually measured at the organism level and used comparatively across species (McGill et al. 2006). On neglecting the intra-species variability which is often small compared to the inter-species variable (Garnier et al. 2001), a trait is a species property (Kleyer et al. 2008). If traits are important in structuring communities, then the composition of local communities should be a non-random sample from the regional species pool (Ozinga et al. 2005a, Shipley et al. 2006). Environmental conditions, such as nutrient availability and soil moisture for plants, can act as filters that alter the probabilities of species to enter a local community according to their trait states (Weiher et al. 1998, Ozinga et al. 2004, Ozinga et al. 2005b, Cornwell and Ackerly 2009). Several empirical studies have shown that species traits are associated with habitat conditions (Townsend

and Hildrew 1994, Townsend et al. 1997, Pöyry et al. 2008) in accordance with the theory that traits in the evolutionary process adapt to the habitat and landscape characteristics in which the species occur (Southwood 1977, Ackerly 2003). Our interest is in finding functional traits from the joint statistical analysis of the three data tables (Fig. 1). We will do so by adding species traits to models that relate species to the environment.

The key questions when modeling species, traits and environments are: (a) how does the expected abundance of species depend on trait and environmental values and (b) which traits and environmental variables best explain the distribution of abundance in space and time and (c) to what extent are traits associated/correlated with environmental variables (Legendre et al. 1997). For modeling different approaches are used. Some authors combined the sites × species table **Y** and species × trait table **Z** in to a sites × trait table which is then related to the sites × environment table **X** by standard statistical methods (Díaz et al. 1992, Sonnier et al. 2010). Legendre et al. (1997) and Dray and Legendre (2008) integrated these two steps in to one, the fourth-corner problem, in which they fill the trait × environment corner that is missing in Fig. 1. The entries of the missing corner table are Pearson correlations between traits and environmental variables, when quantitative, calculated from an inflated table. Dray and Legendre (2008) examined six permutation-based methods to test the statistical significance of the trait-environment relationship, but none of them truly controlled the type I error. The multivariate version of the fourth-corner problem is the RLQ ordination (Dolédec et al. 1996) which has been used for selecting the best traits in plant functional trait analyses (Bernhardt-Römermann et al. 2008). These methods focus on key question (c).

A focus on key question (a) can be found in Shipley et al (2006) and Ozinga et al. (2005a). Shipley et al (2006) used the above mentioned sites × trait table in a novel way as a macroscopic feature of communities to predict species abundance in sites by the maximum entropy principle. The result is a logistic model relating abundances to traits, as used in logistic regression but fitted in a different way (He 2010) and without environmental variables. Ozinga et al. (2005a) started from the multiple logistic regression method and used it to quantify the effect of functional traits in a way that accounts for spatial variation in the composition of the local species pool. Their method assumes that species records within sites are independent (Hosmer and Lemeshow 2000), thus commits pseudo-replication (Hurlbert 1984, Crawley 2002). In applying generalized linear models, such as logistic regression, researchers often ignore the hierarchical structure of the data thereby producing incorrect variance estimates and increasing the likelihood of committing type I error (Wagner and Fortin 2005, Gillies et al. 2006).

To address all three key questions, we develop a dedicated Generalized Linear Mixed Model (GLMM). GLMMs are as very general powerful class of statistical models in ecology and elsewhere (Gelman and Hill 2007, Bolker et al. 2009, Zuur et al. 2009). We introduce our GLMM

as the result of integrating a two-step procedure into one, so obtaining a GLMM with main effects for traits and environmental variables as well as interaction effects between them. The GLMM utilizes species trait data efficiently and overcomes the problem of pseudo-replication (Paterson and Lello 2003). For fitting the model we use the library lme4 (Bates et al. 2011) in the free software package R (R Development Core Team 2011). Other statistical packages with good GLMM facilities include SAS proc glimmix (Stroup 2011) and Genstat (http://www.vsni.co.uk/software/genstat/). In the main text we will use presence-absence abundance data, but the method can be used equally well to count and multinomial data as we show in Appendix S1 in Supplementary Information.

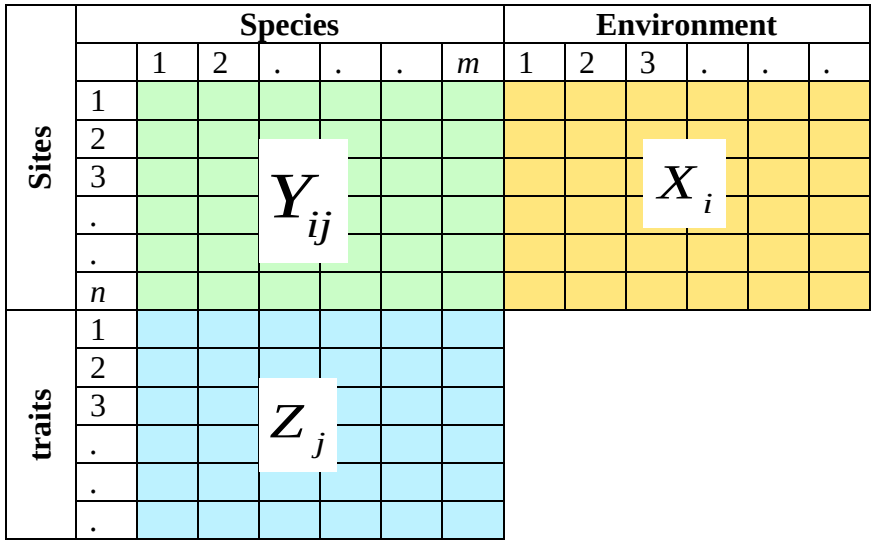

Fig. 1. A table **Y** ($n \times m$) containing the abundances of m species at n sites, a second table **X** ($n \times p$) with measurements of p environmental variables for the n sites, and a third table **Z** ($m \times s$) describing s traits for the m species.

Methods

The data set

We illustrate the method on the basis of the Dune Meadow data (Jongman et al. 1995). This is a small data set of 28 higher plants in 20 sites with five environmental variables and four species traits (Table 1).

The Generalized Linear Mixed Model

In this section, we derive our generalized linear mixed model (GLMM) from a two-step approach. The data we consider is a binary data table $\mathbf{Y} = [y_{ij}]$ recording the presence (1) -absence (0) of m species (columns) in n sites (row), an environmental variable $\mathbf{x} = [x_i]$ with quantitative

Table 1. Abbreviations for environmental variables and traits

Environmental variables	
A1	Thickness of A1 horizon
Moist	Moisture content of the soil
Mag	Grassland management type
Use	Agriculture grassland use
Manure	Quantity of manure applied
Traits: Habitat requirements (Ellenberg indicator values)	
F	Moisture, ranging [1 to 12] (low to high)
R	Soil acidity, ranging [1 to 9]
N	Nitrogen requirement, ranging [1 to 9]
L	Light requirement, ranging [1 to 9]

measurements in the *n* sites, and a quantitative trait **z** = [z_j] with quantitative values for the *m* species. The subscripts *i* and *j* refer to site *i* and species *j*, respectively.

A natural way to study the relationship between a trait and an environmental variable on the basis of species presence-absence data is in two steps, consisting of

1. fitting, for each species separately, a logistic regression of its presence-absence against the environmental variable *x* and
2. regressing parameters retrieved from the *m* logistic regressions on to the trait *z*.

In its simplest form, the first step involves a linear-logistic regression and models the probability of occurrence as a function of the environmental variable. The first stage of two stage approach assumes that

$$\Pr(y_{ij}) = \text{logit}^{-1}(\alpha_j + \beta_j x_i), \qquad i = 1,2,\ldots,n, \;\; j = 1,2,\ldots,m, \tag{1}$$

where Pr(.) is the probability of occurrence of species *j* in site *i*, α_j and β_j are the intercept and slope for j^{th} species with respect to environmental variable *x* and $\text{logit}^{-1}(.) = \exp(.)/(1 + \exp(.))$, the inverse of the logistic function. Extensions of this simple model will be discussed later. This equation can be fitted to the presence-absence data of each species separately, resulting in *m* separate models for the probability of occurrence of the species as a function of the environmental variable *x*. In this model, the relationship of a species with the environment is summarized by the slope β_j. Its sign indicates whether the probability of occurrence increases or decreases with increasing value of *x* and its size how strongly. In its simplest form, the second step involves a (possible weighted) linear regression of the estimated regression slope coefficients {β_j} on to the trait with the model

$$\beta_j = b_0 + b_1 z_j + \varepsilon_{\beta j}, \quad j = 1,2,\ldots,m, \tag{2}$$

with b_0 and b_1 intercept and slope respectively and error $\varepsilon_{\beta j}$, normally distributed with zero mean and variance σ_β^2, *i.e.* $\varepsilon_{\beta j} \sim N(0, \sigma_\beta^2)$. The subscript β is added to the error term to distinguish it from other error terms later on. (The weights are the inverse of the squared standard errors of estimate of $\{\beta_j\}$ in step 1). Another way of expressing Eq. 2 is that the slope coefficient of the species j with trait value z_j is normally distributed with mean $b_0 + b_1 z_j$ and variance σ_β^2, *i.e.*

$$\beta_j \sim N(b_0 + b_1 z_j, \sigma_\beta^2). \tag{3}$$

But, Eqs.1 and 3 together form an example of a generalized linear mixed model (GLMM) and can thus be integrated and estimated simultaneously.

So far the second step only modeled the slopes, because of the particular interest in the trait-environment relationship, but we may also be interested in the influence of the trait on the overall probability of occurrence of a species. The intercept α_j in Eq. 1 plays such a role, in particular when the environmental variable x is centered prior to the analysis, as $\text{logit}^{-1}(\alpha_j)$ is the probability of occurrence at mean x. Analogously to Eq. 2, we could linearly regress the estimated intercepts $\{\alpha_j\}$ on to the trait with the model

$$\alpha_j = a_0 + a_1 z_j + \varepsilon_{\alpha j}, \quad j = 1,2,\ldots,m, \tag{4}$$

with a_0 and a_1 intercept and slope, respectively and $\varepsilon_{\alpha j}$ normally distributed with zero mean and variance σ_α^2. As in Eq. 3 we rewrite this as

$$\alpha_j \sim N(a_0 + a_1 z_j, \sigma_\alpha^2). \tag{5}$$

Eqs.1, 3 and 5 together form another example of a generalized linear mixed model (GLMM). As a GLMM this model still has two shortcomings. First, it assumes that the intercept α_j and slope β_j are independent. This is not very realistic, so we complete the model with a correlation ρ between them. Second, it assumes that the presence-absences of different species at the same site (given their trait values and x_i) are independent. The usual way to introduce correlation among them is with a common site-specific parameter γ_i that is assumed to be normally distributed with mean zero and variance σ_γ^2. With this parameter included, the GLMM equations become

$$\Pr(y_{ij}) = \text{logit}^{-1}(\alpha_j + \beta_j x_i + \gamma_i), \qquad i = 1,2,\ldots,n, \;\; j = 1,2,\ldots,m, \tag{6}$$

$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim N\left(\begin{pmatrix} b_0 + b_1 z_j \\ a_0 + a_1 z_j \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix} \right)$$

$$\gamma_i \sim N(0, \sigma_\gamma^2).$$

This completes our derivation of the GLMM that models the species presence as a function of both the environmental variable x and trait variable z. In the GLMM literature the model is called a

random intercept and random slope model. This GLMM combines both steps of the two-step approach into a single model and avoids pseudo-replication by including site as a random effect.

Testing and interpreting the trait-environment relationship

Here we show that the trait-environment relationship is an interaction term in the model that can be tested for statistical significance using standard software.

By inserting Eqs (4) and (2) in Eq. 6 we obtain

$$\Pr(y_{ij}) = \text{logit}^{-1}\left((a_0 + a_1 z_j + \varepsilon_{\alpha j}) + (b_0 + b_1 z_j + \varepsilon_{\beta j})x_i + \gamma_i\right)$$

$$= \text{logit}^{-1}(a_0 + a_1 z_j + b_0 x + b_1 z_j x + \varepsilon_{\alpha j} + \varepsilon_{\beta j} x_i + \gamma_i) \quad (7)$$

with fixed coefficients in Roman and random coefficients in Greek. This model for the probability of occurrence contains main effects for the trait *z* and the environmental variable *x* and an interaction *z.x* between them. This interaction represents the trait-environment relationship. The model also contains random terms for species ($\varepsilon_{\alpha j}$), sites (γ_i) and the environment-by-sites interaction ($\varepsilon_{\beta j} x_i$). We need to specify all effects and random terms to fit the model to data. With the lme4 library of the software package R the specification of Eq. 7 is

```
M1 <- glmer(y ~ z + x + z:x + (1 + x | species) + (1 | sites),
family=binomial(link="logit"), data)
```

with y, z and x vectors with *nm* elements and species and sites factors with *m* and *n* levels respectively. The terms between parentheses are random, the others are fixed. Library lme4 uses vector notation, *i.e.* y is the matrix $\mathbf{Y} = [y_{ij}]$ written as a vector; the species and site factors code to which species and site each element of y belongs; the value x_i of the environmental value repeated at all *m* elements that code for site *i* and the value z_j of the environmental value repeated at all *n* elements that code for species *j* (Table S1). To test the trait-environment interaction (with null-hypothesis: $b_1 = 0$), we also fit the model without this term by

```
M0 <- glmer(y ~ z + x +(1 + x | species)+(1 | sites),
family=binomial(link="logit"), data)
```

and then compare the two models by an analysis of variance statement anova(M0,M1), resulting in a P-value for the likelihood ratio (LR) test of model M1 against M0.

The estimates of the variance σ_β^2 in model M0 and M1 can be usefully compared to express the contribution of the trait to the inter-species variance in the slope parameter by the coefficient (Grosbois et al. 2009, Lahoz-Monfort et al. 2011)

$$C_\beta = 100\left(1 - \frac{\hat{\sigma}^2_{\beta\,(\text{res})}}{\hat{\sigma}^2_{\beta\,(\text{total})}}\right) \quad (8)$$

Where $\hat{\sigma}^2_{\beta\,(\text{res})} = \hat{\sigma}^2_\beta$ in model M1 and $\hat{\sigma}^2_{\beta\,(\text{total})} = \hat{\sigma}^2_\beta$ in model M0. The rationale is that $\hat{\sigma}^2_\beta$ is the residual variance in Eq. 2, the inter-species variance of the slope parameter after taking account of the trait and therefore denoted as $\sigma^2_{\beta(\text{res})}$. In model M0, b_1 = 0 in Eq. 2, so that σ^2_β represents the total variance denoted by $\sigma^2_{\beta(\text{total})}$.

We investigated the type I error and power of the statistical tests on trait-environment interaction. We simulated 1000 new datasets of the same size and the same environment and trait values as the Dune Meadow data. The data $\{y_{ij}\}$ were simulated using the GLMM model of Eq. 7 with parameters and variance components equal to the estimated ones, i.e. those of model M0 and M1 for the type I error and power calculations, respectively. We did not observe much difference between the test based on the z-statistic and the LR test and report the latter only.

So far, the environmental variable and the species trait were both quantitative. GLMM can also be applied when both are qualitative or when one is quantitative and the other qualitative. A difference is that each class of an environmental factor comes with its own variance component and the trait-environment interaction may consist of more than one regression parameter, but neither difference presents a problem to LR testing and further interpretation. For details see Appendix S2.

Model selection with many environmental variables and traits

The GLMM of Eq. 7 can readily be extended to more traits and environmental variables by including a) main effects for all traits and environmental variables, b) interactions between each trait and each environmental variable, and c) species-dependent random terms for each environmental variable. Conceptually such a model can still be viewed as one with slope coefficients with respect to each of the environmental variables, which are then each (separately) regressed on to the traits. GLMM does a joint fit of such a model. With two environmental variables (x1 and x2) and three traits (z1, z2 and z3), this model that can be specified in lme4 by

```
glmer(y ~ (z1+z2+z3)*(x1+x2)+(1+x1+x2|species)+(1|sites),
family=binomial(link="logit"), data)
```

This model contains trait and environmental variable main effects and their interactions, (correlated) species-dependent random effects for all environmental variables and independent random effects for species and sites.

A natural question is then to select a minimal model that describes the species occurrences data well and, related to this, to select the traits that explain the species response to relevant environmental variables. For an RLQ approach to the latter sees Bernhardt–Römerman et al.

(2008). These questions can be solved by model selection (Diggle et al. 2002, West et al. 2006). The number of candidate models increases exponentially with the number of predictors (traits, environmental variables, interactions and variance components) so an exhaustive search is feasible only for low numbers of predictors. Alternatives are forward and backward selection. Backward selection would start with the model with all terms included. This model may be difficult to fit, due to convergence problems, unless the number of environmental variables is small. For example, we could not fit the full model to the Dune Meadow data using lme4. Therefore, we propose a forward selection approach that starts with the null model with only random effects for species and sites and then adds in each step the environmental variable for which the species-dependent random terms most increases the log-likelihood. So, in the first round the model is that of Eq. 6 with random coefficients α_j and β_j. This process is continued until the increase in log-likelihood is no longer statistically significant as judged on the basis of the LR test. At this first stage, the main effects of traits and environmental variables are not considered because the random species and site effects can already partly take account of them. After this first stage, the choice for the random part of the model is complete.

In the second stage, we consider only the trait-environment interactions of the environmental variables that were selected in the first stage. The reason is that the importance of the trait-environment coefficient (b_1) can only be judged against the unexplained variation in the slope coefficients $\{\beta_j\}$, as can best be seen from Eq. 2. Before interactions can be added we must deal with the associated main effects. A simple approach is to first add the main effects of all environment variables selected in the first stage and the main effects of all traits. This is feasible if the number of traits is smaller than the number of species. In each subsequent step we then search for the trait-environment interaction that most increases the log-likelihood. This process is continued until the increase in log-likelihood is no longer statistically significant. In a final third round we delete sequentially any insignificant main effects that have no associated interaction effect.

The model selection process needs modification when some of the environmental variables or traits are qualitative with more than two classes. Additions may then involve different numbers of parameters (degrees of freedom, df) so that increases in log-likelihood are no longer comparable and must be balanced against the number of degrees of freedom. This is achieved in information criteria such as the Akaike Information Criterion (AIC) (Broman and Speed 2002) which is defined as minus two log-likelihood plus two times df. We must thus look for the model with the lowest AIC value. We use a variant, SigAIC, which multiplies df by $\chi^2_{1(0.05)} = 3.84$ instead of by 2 (Broman and Speed 2002). With SigAIC, the addition of a single parameter to a model will result in a lower SigAIC value if and only if that parameter is significant at the 5% level as judged by the LR test.

Table 2. Parameter estimates with standard error and z-statistic (estimate/standard error) from GLMM with and without traits (M1 and M0).

Model	Fixed effect	Estimate	Standard error	z statistic
M1	intercept	5.045	1.958	
	Moist	-1.931	0.462	-4.18***
	F	-1.056	0.333	-3.17**
	Moist : F	0.322	0.077	4.20***
M0	intercept	-2.077	1.084	
	Moist	-0.029	0.174	-0.17
	F	0.156	0.145	0.28

*P<0.05, **P<0.01,***P<0.001

Results

Relating trait F to moisture

Table 2 illustrates the results for the Dune Meadow data (Jongman et al. 1995) using environmental variable moisture (Moist) and trait the Ellenberg F indicator values (Table 1). As Ellenberg's F ranks the species with respect to the species preference for moisture, we expect a positive relationship. The interaction estimate Moist:F is indeed positive (0.32), showing that species with a high F indicator value have a higher slope coefficient with respect to Moist than species with a low F indicator value. The occurrence probability of species with a high F indicator value thus increases more with Moist than that of species with low F indicator value. The associated z-statistic (estimate/standard error = 4.2) indicates that the interaction is statistically significant (despite the small sample size), so that the true interaction is unlikely to be zero. The LR test (Table S1) confirms that the interaction is highly significant ($P < 0.0001$). In a model with an interaction, the size and sign of main effects depend on the scales of the variables and we explain the interpretation in Appendix S1.

Fig. 2 displays how the fitted occurrence probability depends on Moist for some selected species with and without usage of the trait F in the model (without F, $a_1 = 0$ and $b_1 = 0$ in Eq. 7). The two fitted curves differ more for the species which have few presences. For example, for *Aira praecox* the curve without using trait F is slightly increasing due to one presence at high Moist, whereas it is decreasing with the trait F as *Aira praecox* has a low F value and species with low F typically have a negative regression slope (Fig. S1) and thus a decreasing curve. Similarly, the decrease of the occurrence probability against Moist for *Vicia lathyroides* is stronger with trait usage than without as this species has a low F value. In Appendix S3 we illustrate the advantages of using the GLMM approach over the two-step approach.

GLMM also provides estimates of the variances of the model or their square root; in model M1 (Eq. 6), σ_α is estimated by 2.58, σ_β by 0.55, ρ by -0.87 and σ_γ by 0.37. In model M0, σ_β is estimated as 0.80 showing by Eq. 8 that trait F accounts for 53% of the inter-species variance in the species response to Moisture.

The simulated type I error was somewhat larger than the nominal one (8% and 2% for significance levels 5% and 1%, respectively). By obtaining the critical value not from the chi-square distribution with one degree of freedom, but from an $F_{(1,m-2)}$ distribution decreases the error rates to 7% and 1%. The power of the test was high (99% at a significance level of 5% and 97% at 1%).

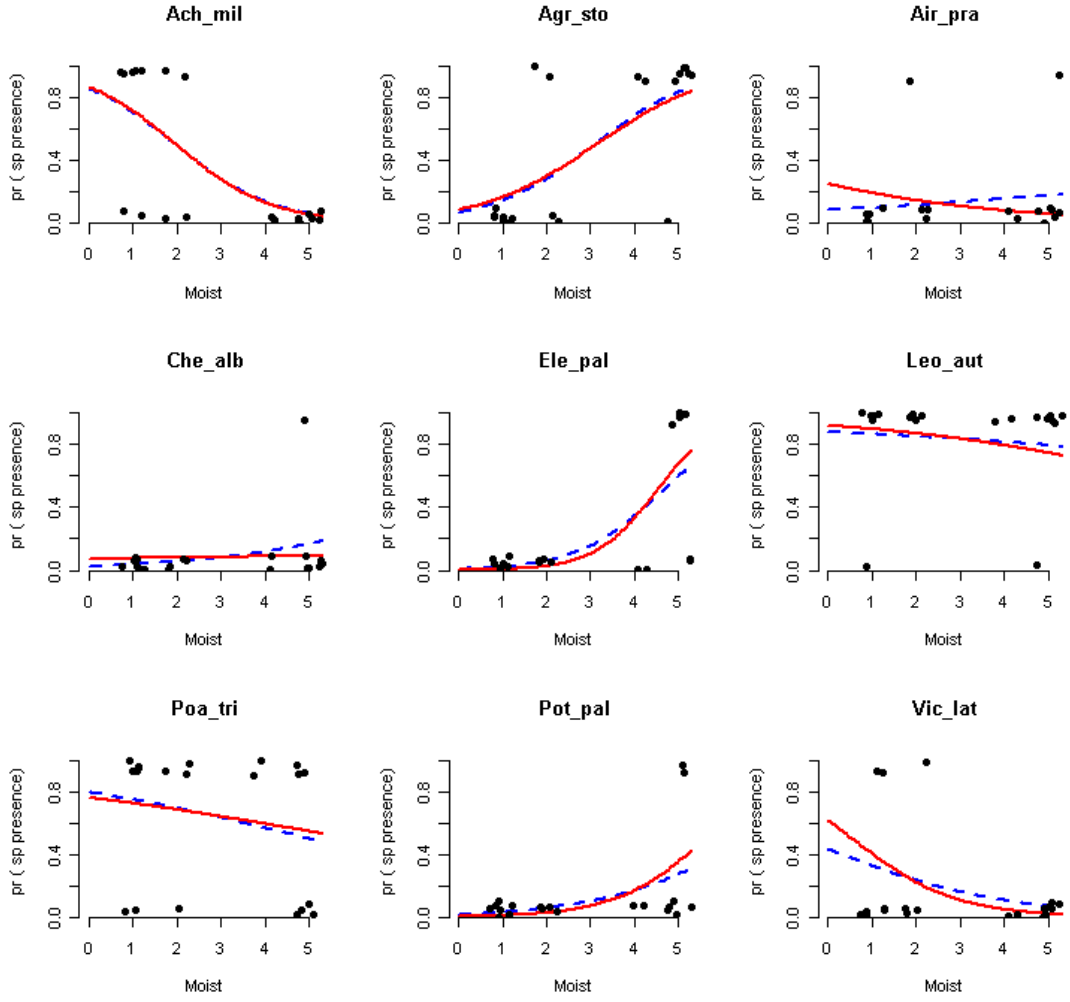

Fig. 2. Occurrence probability against Moisture as fitted by GLMM for nine selected species. Both intercept and slope vary among species and either do (red-solid line) or do not (blue-dashed line) depend on the trait Ellenberg F.
◦**:** jittered presence (1) and absence (0).

Table 3. Three stage forward selection of environmental and trait variables in models explaining species occurrence probability. The variable or interaction (indicated by :) giving the lowest AIC and SigAIC is added in each row (indicated by +). The best model in each stage is indicated in bold.

Stage	Random effects	AIC	SigAIC
1	`(1|species)+(1 | site)`	654.4	650.9
1	`(1+Moist|species)+(1 | site)`	581.2	590.4
1	`(1+ Moist + Manure | species)+(1|site)`	**554.3**	**571.1**
	Fixed effects		
2	`+Moist+Manure+F+R+N+L`	566.9	592.6
2	`+Manure:N`	542.9	570.5
2	`+Moist:F`	533.0	562.4
2	`+Manure:F`	522.9	554.1
2	`+Moist:L`	**520.8**	**554.0**
3	`- R`	**518.9**	**550.1**

Model selection

Table 3 illustrates the selection steps using AIC and SigAIC as selection criteria. In the first stage Moist and Manure are selected. In the second stage the main effects are added. The best interaction to add was that between Manure and N requirement (Manure:N). Three more interactions further decreased both criteria, although the last addition decreased SigAIC only marginally. In the third stage, the insignificant main effect of trait R was deleted.

In the final model traits F and L are significantly positively related to Moisture and traits F and N to Manure (Table S2). The trait main effect N is not significant in the selected model but it is not deleted as it showed significant interaction with Manure. Table 4 shows the variance estimates of the random slopes with respect to Moisture and Manure in models with and without traits, showing that the traits account for 69% and 89% of the inter-species response to Moisture and Manure, respectively.

The correlation between the random intercept and random slope for Moisture is -0.84. It can be reduced in size by centering Moisture. After centering, this correlation was reduced to -0.21 (without essential change in the other statistics). Finally for diagnostic checks we made a Q-Q plot of the random effects to check normality and found that the random effects are reasonable (Fig. S3).

Table 4. Estimated inter-species variance of slope with respect to Moisture and Manure in the final model with (residual variance) and without traits (total variance) and fraction of variance accounted for by traits

Environmental variable	Total variance	Residual variance	Fraction accounted For by traits
Moist	1.05	0.33	0.69
Manure	0.80	0.09	0.89

Discussion

In this paper, we showed how GLMM can be applied for modeling and explaining species response along environmental gradients by species traits. It is based on a sound statistical model that allows, as a standard by-product, questions to be answered about which traits and environmental variables are significantly related and which best explain the species response in a parsimonious model.

GLMM accounts for pseudo-replication and heteroscedastic variance by including sites and species as random factors. Our GLMM approach can be understood as a two-step approach executed at once. In the first step species response is related to the environment and in the second step the (multivariate) outcome of the first step is related to the trait data. The integration of these two steps into one has several advantages: GLMM models directly the variable of interest (occurrence probability, expected abundance), it automatically weighs the different kinds of information for an optimal model fit and standard statistical significance testing and it provides consistent estimates of the between-species variance of (slope) parameters, without introducing unnecessary random variation by replacing the (slope) parameters by their estimates as in the two step approach and it can be applied with small sample size.

In comparison with separate regressions for each species (as in the first step of the two-step approach), the GLMM regression coefficients for each species tend to be pulled inward toward a common value; they are a compromise between the coefficients from a per-species fit and the population average. Such estimates are called shrinkage estimates (Pinheiro and Bates 2000). The shrinkage is particularly evident for the species that have few presences. The estimates for these species lead to abnormally high estimates in the GLM fit (Fig. S1). The pooling of species in the GLMM estimation gives a certain amount of robustness to species with few occurrences in the data.

Our GLMM starts with a logistic linear model (Fig. 2) and is therefore most suitable along short environmental gradients. Such data sets are most common in practice. Moreover, the random component for sites (γ_i) allows for any common non-linearity with as prime example the niche model with equal niche width (Ihm and Van Groenewoud 1984, ter Braak 1988, de Rooij 2007).

One alternative is to convert quantitative environmental variables to qualitative and model how the occurrence probabilities in the newly formed environmental categories depend on the traits, being either quantitative or qualitative. This approach fits in our proposed framework as illustrated in Supplementary Material. Another alternative, adding polynomial terms as random component to the model, is less attractive as it leads to coefficients that lack a clear interpretation.

For the Dune Meadow example data, we found not only the natural associations of moisture and manure with the traits F and N, respectively, but also additional associations of moisture with traits L and manure with trait F. The fourth-corner approach (as implemented in the ade4 package version 1.4-16) with the combined permutation method -the preferred one in Dray and Legendre (2008)- also yields the two natural associations as being statistical significant at the 5% level and two others (F with A1 and R with Manure). A reason for the difference is that the fourth-corner approach tests associations singly. If tested singly, the other two are also significant in a GLMM, but they are insignificant in the multiple variable predictive model obtained after model selection. Also, the fourth-corner approach disregards species absence whereas GLMM takes all data into account.

So far we did neither considered phylogeny, which puts constraints on the way traits may evolve in evolutionary time (Prinzing et al. 2008), nor the spatial configuration of the sites, which set constraints on dispersion (Ozinga et al. 2004, Dray and Legendre 2008). Both aspects can be modeled in a GLMM through additional random terms whose correlation depends on either phylogenic association or spatial distance. These extensions merit further research, also in terms of practical software implementation.

Species traits are likely to have much predictive value for where and when a particular species or group of species appear or disappear. Our model-based approach makes this predictive usage practical and allows the selection of the traits and environmental conditions that matter.

Supporting information

Appendix S1. The GLMM trait model for count and multinomial data

When abundance is a count, Poisson log-linear regression analysis is a commonly used starting point. Poisson log-linear regression is part of the generalized linear model family. The data y_{ij} are assumed to follow a Poisson distribution with mean μ_{ij}

$$y_{ij} \sim \text{Poisson}(\mu_{ij})$$

and the link function is logarithmic function. The analogue of the first part of Eq. 6 in the main text is

$$\text{E}(y_{ij}) = \mu_{ij} = \exp(\alpha_j + \beta_j x_i + \gamma_i)$$

which is usually written as

$$\log(\mu_{ij}) = \alpha_j + \beta_j x_i + \gamma_i$$

The other aspects of the model specification remain the same. In lme4 the GLMM trait model for counts can be fitted by simply replacing "binomial" by "poisson" and "logit" by "log":

```
M1 <- glmer(y ~ z + x + z:x +(1+x|species)+(1|sites),
      family=poisson(link="log"), data)
```

Nothing else changes.

Count data may have a larger variance than assumed by the Poisson distribution. This is called overdispersion and can be detected in the data by introducing using a data-level variance component in the GLMM (Gelman and Hill 2007). The GLMM for overdispersed count data is

$$\log(\mu_{ij}) = \alpha_j + \beta_j x_i + \gamma_i + \varepsilon_{ij}$$

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

The variance component σ_ε^2 measures the amount of overdispersion and can be tested for significance by a LR test. In lme4 we specify

```
data$rows = 1:nrow(data)
M2 <- glmer(y ~ z + x + z:x +(1+x|species)+(1|sites) + (1|rows),
family=binomial(link="poisson"), data)
```

and can test the significance of the overdispersion by
`anova(M1,M2)`.

Multinomial data is data that is count data with a constraint sum so that only the fraction is informative. Abundance data may be modeled as multinomial data as the interest is in the relative abundance only or if the data has been sampled as such, for example, if at each site a pre-specified number of individuals is collected. Multinomial data can be modeled as count data by adding a fixed effect for the factor sites (McCullagh and Nelder 1989)

```
M1 <- glmer(y ~ z + x + z:x + sites + (1+x|species),
      family=poisson(link="log"), data)
```

Unfortunately this specification failed to run in lme4.

References

Gelman A. and Hill J. 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

McCullagh P. and Nelder J.A. 1989. Generalized linear models (second edition). Chapman and Hall, London.

Appendix S2. The GLMM trait model with quantitative and/or qualitative trait and environmental variable

In the main text, the species trait and the environmental variable were both quantitative. The GLMM trait model can also be used to when species trait and environmental variable are both qualitative or when one is quantitative and the other qualitative. Here we illustrate all the combinations with key output and interpretation of the regression coefficients using the Dune Meadow data in vector notation. The R code at the end of this appendix shows for all combinations how to compute the fitted occurrence probability with confidence bands from the estimated regression coefficients and their covariance matrix.

1- Both trait and environmental variable quantitative

As in the main text we consider here the case where both species trait and the environmental variable are quantitative. Now we fit a model, using glmer in the lme4 package, where sp codes for species and site for sites,

```
glmer(y~Moist+Moist:F+(1+Moist|sp)+(1|site), family=binomial, Dune)
```

or to the same effect

```
glmer(y~Moist*F+(1+Moist|sp)+(1|site), family=binomial, Dune)
```

The fixed effects estimates are in Table 1.

Table 1. Fixed effects estimated from GLMM for quantitative environmental variable and quantitative species trait

Fixed effect	Parameter estimate	Standard error	z statistic
(Intercept)	5.045	1.958	2.577**
Moist	-1.931	0.462	-4.18***
F	-1.056	0.333	-3.17**
Moist:F	0.322	0.077	4.20***

The regression equation is ls = 5.045 -1.931 $\times$ Moist – 1.056 $\times$ F +0.322 $\times$ Moist $\times$ F . The result is on logit scale and can be converted to occurrence probability (prob) by

prob = invlogit(ls) = 1/(1+exp(-ls)).

Fig.1 shows the occurrence probability against Moisture for species with different dry preferences (F = 2, 6 and 9) along with 95% confidence bands.

We now return to Table 1. In a model with an interaction, the size and sign of main effects depend on the scales of the variables and may thus be difficult to interpret. Moist runs from 0 to 5 (dry to wet) in the data and F from 2 to 10 (dry to wet preference). In Table 1, the main effect for Moist (-1.93) is negative and significant showing that, if a species would have F = 0, it would decrease in occurrence probability with higher Moist. Such species do not occur in the data; the lowest F is 2. Species with F = 2 still decrease (Fig 1); their slope with respect to Moist is -1.93+2*0.32= -1.29. Species with a high F value, for example F = 9, have a slope of -1.93+9*0.32 = 0.95, indicating that such species are increasing in occurrence probability with higher Moist (Fig 1). The mean F is ~6, giving close to 0 slope (Fig 1), indicating that the occurrence probability does not depend on Moist for such species.

We now turn to the effect of trait F. In model M1, the main effect for F (-1.05) is negative and significant showing that the occurrence probability of the species strongly decreases with increasing F in sites with Moist = 0. Therefore, species that prefer dry conditions (low F) are more like to occur in

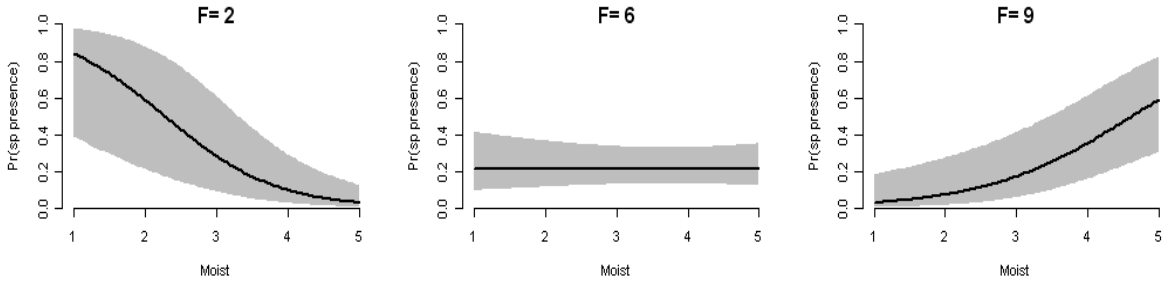

Fig. 1. Occurrence probability of species preferring dry (F=2), intermediate (F=6) and wet (F=9) conditions in relation to moisture in the dune meadow from a GLMM model along with 95% confidence bands.

dry meadows than species that prefer wet conditions (high F). In sites with Moist = 5, the slope with respect to F is -1.05+5*0.322 = 0.56, showing that the occurrence probability of the species increases with increasing F in sites with high Moist.

2- Quantitative trait and qualitative environmental variable

We consider the case where the environmental variable is qualitative and species trait is quantitative. This yields separate regression lines for each category of the environmental variable (Fig. 1). In our example Moist is a qualitative explanatory variable (i.e., a factor), with two categories: Moistdry and Moistwet, depending on whether moisture smaller than 3.5 (Moistdry) or higher (Moistwet).

```
data$moist<-factor(data$Moist)
levels(data$moist)<-list(dry=c(1,2,3),wet=c(4,5))
```

Now we fit a model, using glmer in the lme4 package, with the same type of statement as before

```
glmer(y~moist*F+(1+moist|sp)+(1|site),family=binomial, Dune)
```

The fixed effects estimates are in Table 2.

Table 2. Fixed effects estimated from GLMM for qualitative environmental variable and quantitative species trait

Fixed effect	Parameter estimate	Standard error	z statistic
(Intercept)	2.135	1.304	1.637
moistwet	-6.473	1.476	-4.386***
F	-0.553	0.221	-2.503 *
moistwet:F	1.061	0.242	4.386 ***

The regression equation for Moistdry (the first level of the factor moist) is on logit scale is straightforward

$$\text{Moistdry} \rightarrow 2.134 - 0.553 \times \text{F}$$

The regression equation for Moistwet can be obtained as follows. The intercept for Moistwet can be found by adding the coefficients for intercept and Moistwet and the slope for Moistwet with respect to F by adding the coefficients for F and moistwet:F . The regression equation for Moistwet becomes on logit scale

$$\text{moistwet} \rightarrow (2.135 - 6.473) + (-0.553 + 1.0610) \times \text{F} = -4.338 + 0.508 \times \text{F}$$

Both equations can be converted to occurrence probability curves with confidence bands (Fig. 2).

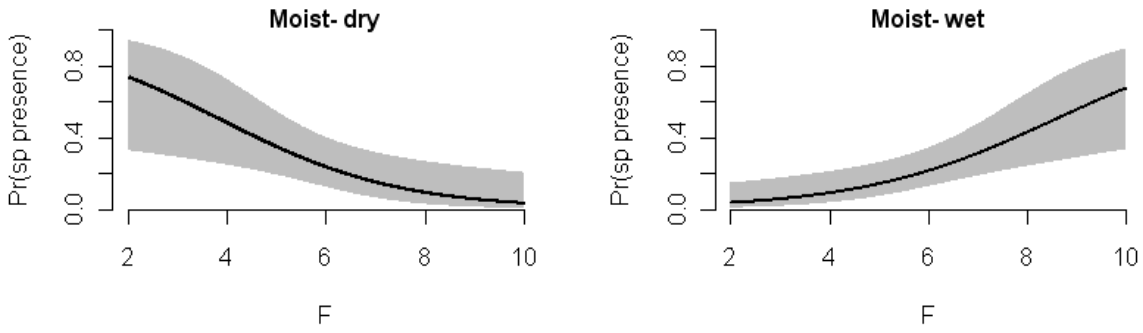

Fig. 2. Occurrence probability of a species in dry and wet meadows, with 95% confidence band, in relation to the species trait F from a GLMM model where the environmental variable is a factor with two categories.

Fig. 2 shows that in dry meadows the probability of occurrence of species decreases with increasing Ellenberg indicator F, whereas in wet meadows the probability of occurrence of species increases with increasing Ellenberg indicator F. The two regression lines are crossing, so showing interaction between Moist and F. In Table 2 the interaction is represented by one regression coefficient (1.0610) which is high significant.

A trick to immediately obtain the regression model for each meadow category is to make a slight modification in the model specification:

```
glmer(y~0+moist+moist:F+(0+moist|sp)+(1|site),family=binomial, Dune)
```

The results for this model specification are displayed in Table 3.

Table 3. Fixed effects estimated from GLMM for qualitative environmental variable and quantitative species trait

Fixed effect	Parameter estimate	Standard error	z statistic
moistdry	2.135	1.304	1.637
moistwet	-4.338	1.120	-3.874***
moistdry:F	-0.553	0.221	-2.503 *
moistwet:F	0.508	0.173	2.943 **

Table 3 contains directly the coefficients of the regression equations for Moist-dry and Moist-wet equation on logit scale:

Moistdry $\rightarrow$ $2.134 - 0.553 \times$ F
Moistwet $\rightarrow$ $-4.338 + 0.508 \times$ F

In Table 3 there seem two interaction terms, but the real interaction is the difference between the two. Table 2 and Table 3 use different parameterizations of the same model. In either case, a likelihood ratio (LR) test of the interaction is obtained by comparison with the model

```
glmer(y~moist+F+(1+moist|sp)+(1|site), family=binomial, Dune)
```

using the anova() statement.

3- Qualitative trait and quantitative environmental variable

We consider the case where the trait is qualitative and the environmental variables is quantitative. This yields separate regression lines for each trait category (Fig. 3). In our example F is a qualitative explanatory variable (i.e., a factor), with two categories: dry and wet, depending on whether F smaller than 5.5 (dry) or higher (wet).

```
data$F= factor(data$F)
levels(data0$F)<-list(dry=c(2,4,5),wet=c(6,7,8,9,10))  # 2 levels
```

Now the model specification using glmer in the lme4 package

```
glmer(y~Moist*F+(1+Moist|sp)+(1|site), family=binomial, Dune)
```

The fixed effects for the above model are given in Table 4.

Table 4. Fixed effects estimated from GLMM for qualitative trait and quantitative environmental variable

Fixed effect	Parameter estimate	Standard error	z statistic
(Intercept)	0.760	0.741	1.026
Moist	-0.529	0.200	-2.643 **
Fwet	-3.662	1.066	-3.436 ***
Moist:Fwet	0.952	0.278	3.421***

The regression equation for F-dry (the first level of the factor F) is on logit scale is straightforward

$$\text{F-dry} \rightarrow 0.760 - 0.529 \times \text{Moist}$$

The regression equation for F-wet can be obtained as follows. The intercept for F-wet can be found by adding the coefficients for intercept and F-wet and the slope for F-wet with respect to moisture by adding the coefficients for Moist and Moist:F-wet . The regression equation for Moistwet becomes on logit scale

$$\text{F-wet} \rightarrow (0.760 - 3.662) + (-0.529+0.952) \times \text{Moist} = -2.902 + 0.423 \times \text{Moist}$$

Both equations can be converted to occurrence probability curves (Fig. 3).

The probability of occurrence of species that prefer dry meadows decreases with increasing moisture, whereas the probability of occurrence of species that prefer wet meadows increase with increasing moisture (Fig. 3).

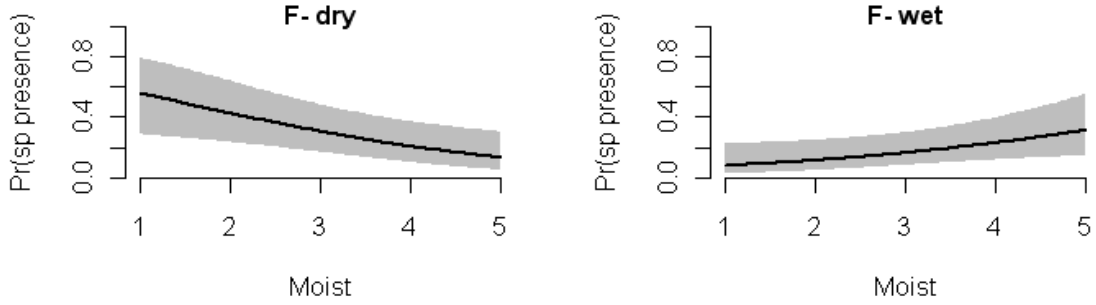

Fig. 3. Occurrence probability of species preferring dry (F<3.5) and wet (F>3.5) conditions, with 95% confidence band, in relation to moisture in the dune meadow from a GLMM model where the trait is a factor with two categories

4- Both trait and environmental variable qualitative

We consider the case when both trait and environmental variable are qualitative. This yields occurrence probabilities in each class of the cross-classification of trait and environment. In our example, the species trait F and environmental variable moist are each classified into two categories as above. The model specification in R is:

```
glmer(y ~moist*F +(1+moist|sp)+(1|site), family=binomial, Dune)
```

Table 5. Fixed effects estimated from GLMM for quantitative environmental variable and quantitative species trait

Fixed effect	Parameter estimate	Standard error	z statistic
(Intercept)	0.061	0.475	0.127
moistwet	-1.893	0.634	-2.985 **
Fwet	-2.233	0.688	-3.247 **
moistwet:Fwet	3.235	0.878	3.685***

From the coefficients in Table 5 we need to construct the occurrence probabilities in each class. The reference class is the first level of moist (dry) and the first level of F (dry), moistdry-Fdry; we have on the logit-scale

for class moistwet-Fwet $\rightarrow$ 0.061
for class moistwet-Fdry $\rightarrow$ 0.061-1.893 = -1.832
for class moistdry-Fwet $\rightarrow$ 0.061-2.233 = -2.172
for class moiswet- Fwet $\rightarrow$ 0.061-1.893-2.233+3.235 = -0.83

The occurrence probabilities are the inverse logit of these values, for example invlogit(0.061) = 1/(1+exp(-0.061))= 0.515.

Table 5 shows all four probabilities and 95% confidence limits. In dry meadows the probability of occurrence for species that prefer dry condition is higher than for species that prefer wet condition and in wet meadows the probability of occurrence for species that prefer dry condition is lower than for species that prefer wet condition.

Table 6. Probability of occurrence and in parentheses are the corresponding confidence limits of species in dry and wet meadows according to their preference for dry conditions (F<3.5) or wet conditions (F>3.5).

		F	
		dry	wet
Moist	dry	**0.515** (0.29, 0.72)	**0.102** (0.04, 0.24)
	wet	**0.138** (0.05, 0.31)	**0.303** (0.15, 0.53)

5- R code

The R code for glmm-plot-conf-int.r.

```
rm(list=ls(all=TRUE))
library(lme4)
library(arm)
Dune=read.table("Dune.txt", header=TRUE,sep=" ")
colnames(Dune)

glPredict <- function(fm1, newdat, conf = 95) {
  # Predicts occurrence probability with confidence limits from an lmer object
  # at the points provided as rows of newdat
  # fm1 = lmer  object
  # newdat =data frame with values for predictors for which prediciton must be
  #made
  # confidence value (in %)
  # for related code see package ez
  # Value:
  #  y, lo, hi = prediction with confidence limits on link scale
  #  p, plow, phigh =occurrence probability with confidence limits

  frac = 1 - (100-conf)/200
  mm = model.matrix(terms(fm1),newdat)
  y = mm %*% fixef(fm1)  # prediction on link scale
  Var <- Matrix::diag(mm %*% tcrossprod(vcov(fm1),mm)) # variance on link scale
  lo = y-qnorm(frac)*sqrt(Var)
  hi = y+qnorm(frac)*sqrt(Var)
  newdat$y = y
  newdat <- data.frame(newdat, ylo = lo, yhi = hi,
    p = invlogit(y), plow = invlogit(lo), phigh = invlogit(hi))
  newdat
}

#####################################################################
# Table 1 #quant env; quant trait
#####################################################################

fm1 = lmer (y ~ Moist *F +(1 + Moist | sp)+(1|site),
            family=binomial,data=Dune)

# for confidence limits
newdat <- expand.grid( Moist=seq(1,5,length.out=100), F = c(2,6,9 ), y =0)
newdat <- glPredict(fm1, newdat)
names(newdat)
# plotting
par(bty="n")
par(mfrow=c(1,3))
for ( j in c(2,6,9)){

  data.f<- subset( newdat , F %in% j)
  x<- data.f$Moist
  plot(0,0,ylim=c(0,1),xlim=range(x),ylab="Pr(sp presence)" ,xlab="Moist" ,
      yaxs="i" , main="",type="n")
  mtext(paste("F=",j ),   font= 2, col= "black" )
  polygon(c(x, rev(x)),c(data.f$phigh, rev(data.f$plow)),col='gray',border =
FALSE)
  points(x, data.f$p, type='l',lwd=2.5)
}
#####################################################################
# Table 2 #factor env; quant trait
#####################################################################
Dune$moist= factor(Dune$Moist)
levels(Dune$moist)<-list(dry=c(1,2),wet=c(4,5))

print(fm2<-lmer(y~ moist*F+(1+moist|sp)+(1|site)
                , family=binomial, Dune),corr=FALSE)
```

```
newdat <- expand.grid(moist =c("dry","wet"),F=seq(2,10,length.out=1000),y = 0)

newdat <- glPredict(fm2, newdat)

par(mfrow=c(1,2))
for ( j in c("dry","wet")){
    data.f<- subset( newdat , moist %in% j)

    x<- data.f$F
    plot(0,0,ylim=c(0,1),xlim=range(x),ylab="Pr(sp presence)" ,xlab="F" ,
          yaxs="i" , main="",type="n")
   # title(paste("F=",j ), cex.main = 1.2,   font.main= 2, col.main= "black")
    mtext(paste("Moist-",j ),   font= 2, col= "black" )
    polygon(c(x, rev(x)),c(data.f$phigh,rev(data.f$plow)),col='gray',
border=FALSE)
    points(x, data.f$p, type='l',lwd=2.5)
}
# the alternative parametrization
print(fm2.B<-lmer(y~0+moist+moist:F+(0+moist|sp)+(1|site)
                  , family=binomial, Dune),corr=FALSE)
newdat.B <- glPredict(fm2.B, newdat)
all.equal(newdat.B,newdat)

fm0<-lmer(y~moist+ F+(1+moist|sp)+(1|site)
                  , family=binomial, Dune)
anova(fm0,fm2)
anova(fm0,fm2.B)
#################################################################
# Table 3 #quan env; factor trait
#################################################################
Dune$F= factor(Dune$F)
levels(Dune$F)<-list(dry=c(2,4,5),wet=c(6,7,8,9,10))   # 2 levels


print(fm3<-lmer(y~Moist*F+(1+Moist|sp)+(1|site)
                  , family=binomial, Dune),corr=FALSE)

newdat <- expand.grid( Moist=seq(1,5,length.out=100),F=c("dry","wet"), y =0)
newdat <- glPredict(fm3, newdat)

par(mfrow=c(1,2))
for ( j in c("dry","wet")){
    data.f<- subset( newdat , F %in% j)
    x<- data.f$Moist
    plot(0,0,ylim=c(0,1),xlim=range(x),ylab="Pr(sp presence)" ,xlab="Moist" ,
          yaxs="i" , main="",type="n")
    mtext(paste("F-",j ),   font= 2, col= "black" )
    polygon(c(x, rev(x)),c(data.f$phigh,rev(data.f$plow)),col='gray',border=
FALSE)
    points(x, data.f$p, type='l',lwd=2.5)
}

#################################################################
# Table 4 #factor env; Factor trait
#################################################################

Dune$moist= factor(Dune$Moist)
levels(Dune$moist)<-list(dry=c(1,2),wet=c(4,5))
#factor env; quant trait
print(fm4<-lmer(y~moist*F+(1+moist|sp)+(1|site)
                  , family=binomial, Dune),corr=FALSE)

newdat <- expand.grid( moist=c("dry","wet"), F = c("A","B" ), y = 0  )
newdat <- glPredict(fm4, newdat)
newdat
newdat[, -(3:5)]

#end
```

Appendix S3 Comparison of GLMM with the two-step approach

Fig. S1 shows the estimated slopes $\{\beta_j\}$ against trait F for the GLMM and the two-step approach. The slopes of the two-step approach are extremely large in absolute value (>3) for species with few occurrences. In the GLMM approach the slopes are shrunken towards the common regression line; the vertical deviations from the line are summarized by the parameter σ_β = 0.55 in equation (6) of the main text.

The dashed regression line of the two-step approach is fitted by weighted least-squares and shows a weaker relationship than that of GLMM. A small simulation study was done to see whether that was incidental. In the 99% of the 1000 simulated data sets of the power study, the coefficient b_1 estimated by GLMM was greater than that in the two-step approach. It was also closer to the true coefficient as judged from the root mean squared error (0.071 compared to 0.119). In GLMM, the standard deviation across simulated data sets (0.066) was close to the standard error of estimate reported in Table 2 (0.077), showing that this standard error of estimate is valid in this data.

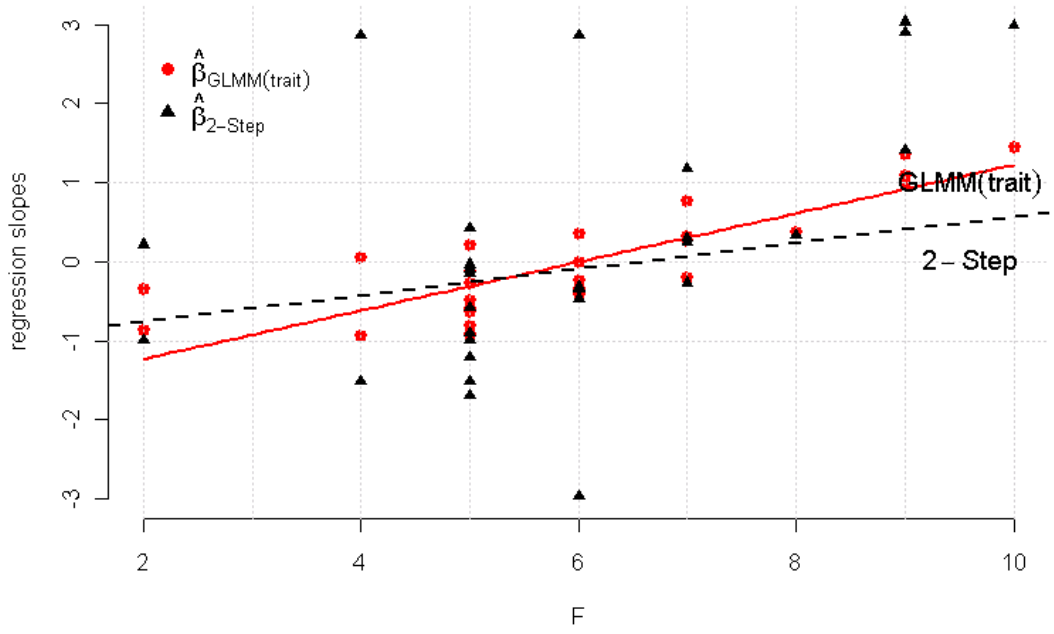

Fig. S1. Estimated regression slopes (β_j) of species versus trait F with fitted regression line for GLMM (circles with red solid line) and the 2-step approach (triangles with black dashed line).

Table S1. Comparison of models with (M1) and without (M0) trait-environment interaction by anova(M0,M1). df = degrees, logLik = loglikelihood, Chi-sq = difference in logLik, Chi df = difference in df, Pr(>Chisq) = P-value.

Model	df	logLik	Chi-sq	Chi df	Pr(>Chisq)
M0	7	-285.28			
M1	8	-277.48	15.61	1	7.794e-05***

Table S2. Parameter estimates with standard error and z-statistic from GLMM in the final model (after model selection):
y ~ Moist × (F + L) + Manure × (F + N) + (1 + Moist + Manure | sp) + (1|site).

Effects	Fixed effect	Parameter estimate	Standard error	z statistic
Main	Moist	-5.50	1.44	-3.81***
	Manure	-3.82	0.73	-5.20 ***
	F	-1.75	0.40	-4.36*
	L	-1.80	0.72	-2.52***
	N	-0.23	0.18	-1.23
Interactions	Moist : F	0.44	0.09	4.90***
	Moist : L	0.38	0.18	2.15*
	Manure : F	0.29	0.08	3.56***
	Manure : N	0.46	0.07	6.01***

*P<0.05, **P<0.01,***P<0.001

Chapter 3

A unimodal species response model relating traits to environment with application to phytoplankton communities

Tahira Jamil, Carla Kruk, Cajo J.F. ter Braak

Abstract

Niche theory proclaims that species response to environmental gradients is unimodal. For presence-absence data, the simplest unimodal (non-negative) species response curve is the Gaussian logistic response curve with three parameters that characterize the niche: optimum (niche centre), tolerance (niche width) and maximum (expected occurrence at the centre). Niches of species differ between species and species are assumed to be evolutionary adapted. In this paper we attempt to explain the observed niche differences by the differences in traits of the species. To this aim, we propose the trait-modulated Gaussian logistic model in which the niche parameters are made linearly dependent on species traits. The model is fitted to data in the Bayesian framework using OpenBUGS (Bayesian inference Using Gibbs Sampling).

A Bayesian variable selection method is used to identify which species traits and environmental variables best explain the species data through the trait-modulated Gaussian logistic model. The approach is extended to find the best linear combination of environmental variables.

The methods are illustrated using phytoplankton community data of 203 lakes located within four climate zones and associated measurements on 11 environmental variables and six morphological species traits of 60 species. Chlorophyll-a is found to be the best environmental variable, followed by temperature. Chlorophyll-a and temperature are also the most important contributors to the best linear combination of environmental variables with opposite signs of their coefficients. About 25% of the variance in the niche centres with respect to chlorophyll-a could be accounted for by the traits, whereas niche width and maximum could not be predicted. Volume, mucilage and flagella are found to be the most important traits to explain the niche differences.

Key-words: Niche theory; environmental gradient; trait-environment relationship; Gaussian logistic mode; nonlinear mixed model; species traits

Introduction

Phytoplankton is a diverse group of microscopic photosynthesizing algae and cyanobacteria with a short life span of few days. Phytoplankton is fundamental for maintaining global biogeochemical cycles and trophic webs of pelagic ecosystems (Follows et al. 2007), and their excessive growth is one of the main concerning aquatic quality problems (Falkowski et al. 2003). The identification of the main biotic and abiotic factors controlling phytoplankton in lakes is essential for management of ecosystem (Peretyatko et al. 2007). As phytoplankton community composition impacts the functioning of aquatic ecosystems and thereby indirectly global climate, it is important to understand what factors regulate phytoplankton community assembly and dynamics.

The construction of species habitat templates (species niches) is necessary to predict community structure changes with changing environments. Species can be characterized by a large number of quantitative and qualitative traits. Species with particular traits will be able to growth under particular environmental conditions and the species habitat template should combine both.

The idea of matching species to habitats templates started early with Tansley (1939) and Pearsall (1950) and was well developed by Grime (1977) for plants. The conceptual basis are credited to Southwood (1977, 1988). Further development was by Keddy (1992) to predict community organization from species pool and species traits, linking traits with environmental conditions. See also the work by Rice et al. (1983) in bird ecology, by Bayley and Li (1992) in fish size and migration, hydrology, by Townsend and Hilderw (1994) in stream ecology, by Wiens (1991) in shrub-desert avifauna and by Statzner et al. (1994) for plants and animals in rivers.

Traits included in the habitat template should be functional, directly or indirectly related to fitness and easy to estimate for any species and organism (Violle et al. 2007, Violle and Jiang 2009). Phytoplankton is a good model to accomplish this objective. Phytoplankton organisms are small and reach high abundances with high growth rates, this enable the rapid track of environmental changes and their study at human scales (Litchman and Klausmeier 2008). Further, functional traits based on morphology which are easy to estimate for any organism (Kruk et al. 2010) and predictable form environmental variables (Kruk et al. 2011). Thus phytoplankton is an excellent model for the construction of species habitat templates.

Following seminal works by Southwood (1977) and Townsend and Hildrew (1994), trait-based approaches have been increasingly applied to explain and predict response of phytoplankton species to environmental conditions. Habitat templates have been built up for phytoplankton for different species, combining traits and environmental gradients by (Margalef 1978, Reynolds 1988, Reynolds et al. 2002) concerning as the main axes growth abilities, resources acquisition and

evasion of loss processes. Formalizations for the construction of habitat templates for phytoplankton have been done mainly by Reynolds combining species preferences and tolerances (Reynolds 1987, Reynolds 1998). However, these conceptual models are difficult to apply to any species in any conditions. The studies that cluster the species based on their functional trait and then summarize their response to environmental change have also been applied by many (Weithoff 2003, Follows et al. 2007, Litchman and Klausmeier 2008). Results from all these studies revealed that traits could offer new insights into phytoplankton ecology.

Like all aquatic organisms, phytoplankton species have preferred environmental conditions in which they can survive and reproduce optimally. Each species is therefore largely confined to a specific interval along an environmental variable. The value most preferred by a species was termed its "indicator value" or optimum. This concept can be extended from one environmental variable to many. Each species is thus presumed to occur in a characteristic, limited range of the multi-dimensional habitat space, called its ecological niche, and within this niche, each species tends to be most abundant around a specific environmental optimum (Green 1971). Therefore, the distribution of species along any environmental gradient is usually unimodal, with the maximum at some ecological optimum.

The simplest unimodal (non-negative) species response curve is the Gaussian response curve. It is symmetric and bell-shaped with three interpretable parameters: the optimum, height of the response and tolerance or width of the curve (Jongman et al. 1995, Oksanen and Minchin 2002). The model can easily be extended to more than a single environmental variable. The model can be fitted by nonlinear regression, but it is easier to first reparametrize it as a generalized linear model (GLM) with a second order polynomial in the environmental variables and then fit it to data by any of the statistical packages that can handle GLMs (ter Braak and Looman 1986, Oksanen et al. 2001). The data can be presence-absence, counts or biomass and for each of these data type there is an appropriate GLM.

Chapter 2 developed a statistical approach to related species traits to environment using an extension of GLM, namely the generalized linear mixed model (GLMM). It uses the environmental variables linearly, so it is unclear whether the model is of any use when data come from an ecosystem with niche structure, i.e. from a unimodal system. In their approach the regression parameters (intercept and slope in a linear model) are made dependent on the species traits. We might try to add squared environmental variables to the model as in the basic analysis of the Gaussian response model. However, regression parameters of linear terms and the squared terms have no intuitive meaning and no ecological interpretation. Moreover, the meaning of the parameter of the linear term depends on the value of that of the squared term and also on the scale used for the environmental variable. It appears therefore rather useless ecologically to make these parameters dependent on the species traits. By contrast, the optimum, the tolerance and the

maximum are interpretable parameters and we would like to model these in terms of the species traits.

As in Chapter 2 we could have attempted a two-step approach by first deriving estimates of the optimum, tolerance and maximum for each species separately by GLM and then regressing these in turn on to the species traits. However, the estimates can be quite variable, in particular for species with low numbers of occurrence. Therefore we propose in this paper an integrated approach. With this in mind, the aim of this paper is to relate species traits to the environment via statistical models that explicitly acknowledge the concept of the ecological niche, i.e. models that are unimodal in terms of the environmental variables.

The Gaussian logistic model (ter Braak and Looman 1986) with linear trait submodels for the parameters, that we propose, cannot be fitted easily with the available (generalized) nonlinear mixed model software. Instead, we take a Bayesian approach and fit the model using OpenBUGS (Bayesian inference Using Gibbs Sampling) (Sturtz et al.)

Crucial to the aim is the identification of those traits (covariates) responsible for explaining the variation in response curve parameters (optimum, tolerance, maximum). The problem is akin to the familiar model selection problem in regression where we try to explain a response variable by a number of explanatory variables (whether continuous or discrete factors). The challenge is to select a small subset of the trait variables that explain a large fraction of the variation in the response parameters. For covariate selection we use the approach of George and McCulloch (1993) extended in Yuan and Lin (2005). The same approach is also used to find the linear combination of environmental variables that best explains the species data through trait modulated Gaussian logistic response curves. The methods are illustrated using phytoplankton communities data. The data has 60 species observed at 203 sites, 11 environmental covariates and 6 trait covariates.

The structure of the paper is as follows. We first give a brief introduction to unimodal response curve, Bayesian theory and its implementation using MCMC algorithms in OpenBUGS. We then present a case study showing how Bayesian variable selection method can select the important environmental variables and traits, where traits are functions of parameters of unimodal response curve. After presenting the results we discuss and interpret the results. Finally we conclude with the implications of this approach and the future extension of our research.

Model

Unimodal response curve

In this section, we propose a trait-modulated Gaussian logistic model. The data we consider here is a binary data table $\mathbf{Y} = [y_{ij}]$ recording the presence (1) -absence (0) of m species (columns) in n

sites (row), an environmental variable x_i (i = 1,…, n) with quantitative measurements in the n sites, and K quantitative or binary traits $\{\mathbf{z}_k , k =1,\ldots,K\}$, with $\mathbf{z}_k =[z_{jk}]$ (j=1,…,m) containing the values of the k^{th} trait for the m species. The subscripts i, j and k refer to site i, species j, and trait k, respectively. Later on we consider the case with multiple environmental variables. We start with the Gaussian logistic model (ter Braak and Looman 1986) with an extra random term for sites (Eq. 1). This term is added to account for the fact that species observed at the same site are likely to correlated in occurrence, even after having taken account of the environmental (and trait) information. The model is phrased in terms of the logit of the probability of occurrence p_{ij} = E(y_{ij}), the expected value of the observation y_{ij}, given the model,

$$\text{logit}(p_{ij}) = a_j - \frac{(x_i - opt_j)^2}{2tol_j{}^2} + \gamma_i^{site} \quad (1)$$

with x_i a quantitative known environmental variable, a_j is a coefficient related to maximum probability, opt_j is the species optimum, tol_j is the tolerance of species, and $\gamma_i^{site} \sim N(0, \sigma_{site}^2)$ the random site effect with variance σ_{site}^2. Recall that logit(.)=log(./(1- .) with inverse 1/(1+exp(-.)). This model has thus a logistic form, and the model parameters opt and tol occur nonlinearly in the model function. The optimum on the gradient gives the location where the maximum probability of occurrence is attained and the tolerance gives the width of the response (ter Braak and Looman 1986).

In the trait-modulated Gaussian logistic model, the parameters a, opt and tol are modulated by the K traits according to the linear submodels

$$a_j = \alpha_0^a + \sum_{k=1}^K \beta_k^a z_{jk} + e_j^a, \quad (2)$$

$$opt_j = \alpha_0^{opt} + \sum_{k=1}^K \beta_k^{opt} z_{jk} + e_j^{opt}, \quad (3)$$

$$tol_j = \alpha_0^{tol} + \sum_{k=1}^K \beta_k^{tol} z_{jk} + e_j^{tol}, \quad (4)$$

with intercepts indicated by α_0 with a superscript for the corresponding parameter and similarly slopes by β_k with k for the associated trait. The error terms in these submodels are $e_j^a \sim N(0, \sigma_a^2)$, $e_j^{opt} \sim N(0, \sigma_{opt}^2)$ and $e_j^{tol} \sim N(0, \sigma_{tol}^2)$ and are usually called the random effects when these equation are inserted in Eq. 1. The resulting model is a nonlinear mixed model, where both fixed and random effects enter nonlinearly. We implemented the model in OpenBUGS and fitted it to phytoplankton community data. OpenBUGS uses Markov Chain Monte Carlo (MCMC), in particular Gibbs sampling, to generate a sample from the posterior distribution.

Statistics for assessing contribution of traits variables

After fitting the model to data, the contribution of individual traits to the model can partly be assessed by the (standardized) size of their slope parameters $\{\beta_k\}$ in Eqs. 2-4. In line with the usual definition of percentage variance explained in a model with multiple predictors, we measure the joint contribution of the *K* traits to the model for the optimum by (Grosbois *et al.*, 2009, Lahoz-Monfort *et al.*, 2011)

$$C_{opt} = 100(1 - \frac{\hat{\sigma}^2_{opt\,(res)}}{\hat{\sigma}^2_{opt\,(total)}}) \ , \qquad (5)$$

where $\hat{\sigma}^2_{opt\,(res)}$ is the estimated variance in the model of Eqs. 1-4 and $\hat{\sigma}^2_{opt\,(total)}$ that in the model with all $\beta_k^{opt} = 0$, for $k = 1,\ldots, K$. In Eq. 5 we compare the variance of the optimum in the model with and without traits (Chapter 2). Analogous definitions of percentage variance explained can made for the tolerance and the maximum. The variances are estimated by the posterior median.

It is worth pointing out that including traits in the model does not constrain the optimum (or tolerance or maximum), such as in constrained ordination (ter Braak and Verdonschot 1995). The reason is that Eqs. 2–4 include a random term, such as e_j^{opt}, whereas such random term is not included in constrained ordination. In our model, including traits attempts to shift unexplained variance, such as $\hat{\sigma}^2_{opt\,(total)}$ as much as possible to the fixed effects of a trait, thereby reducing the unexplained variance to $\hat{\sigma}^2_{opt\,(res)}$. We therefore do not expect much change in the variance explained on the level of the species data $\{y_{ij}\}$.

Bayesian Variable Selection

In data sets with many potential predictors, choosing an appropriate subset of traits and/or environmental variables is a challenging and important task. Here we use the Bayesian variable selection (BVS) approach of Yuan and Lin (2005), the empirical Bayes estimator of which is closely related to the LASSO estimator. The model analyzed here is the unimodal response curve and parameters of the curve have a regression relation with a number of predictors. We apply variable selection to this regression relation within the full model. Here the variable selection is carried out to obtain a parsimonious model with fewer variables. The variable selection is part of larger model. The underlying notion is that most of the traits are expected to have no or only weak effects on the optimum, tolerance and maximum.

Bayesian variable selection can be influenced by the prior. In principle there is considerable flexibility in the priors that could be used. Several Bayesian variable selection methods have been developed in recent years (George and McCulloch 1993, Green 1995, Kuo and Mallick 1998,

Brown et al. 1998, Yuan and Lin 2005). For details and a review of Bayesian model selection methods see O'Hara and Sillanpää (2009).

To keep the presentation simple, assume that the task is to explain an outcome ϕ_j for species *j* (*j* = 1,…,*M*) using *K* trait covariates with values Z_k ; *k*= 1,…, *K*. Naturally, these covariates may be continuous or discrete variables. Given a vector of regression parameters $\theta = (\theta_k)$ of size *K*, the response is modelled as a linear combination of the explanatory variables z_{jk}:

$$\phi_j = \mu + \sum_{k=1}^{K} \theta_k Z_{jk} + e_j \qquad (6)$$

Here μ is the intercept and $e_j \sim N(0, \sigma_\phi^2)$ are the errors. The data are usually sufficiently informative to estimate the overall mean μ and the variance σ_ϕ^2 (the variation in response model parameter). Thus, we can use any reasonably noninformative prior distributions for these parameters. We used uniform priors for μ and σ_φ, *i.e.* $\pi(\mu) \sim 1$ and $\pi(\sigma_\varphi) \sim 1$.

We could assume a normal prior for θ_k. But we expect that most of the θ coefficients are expected to be zero or close to zero. To incorporate this prior knowledge into our analysis, therefore, we can set up a "slab and spike" prior (Miller 2002), with a spike (the probability mass) either exactly at or around zero, and a flat slab elsewhere. By introducing the latent variable γ_j= 0 or 1, we adopt the hierarchical Bayes framework of Yuan and Lin (2005) by assuming a mixture prior for θ_j

$$\theta_j|\gamma_j = (1-\gamma_j)\delta(0) + \gamma_j DE(0,\tau) \qquad j = 1, \ldots, p\,, \qquad (7)$$

where $DE(0,\tau)$ is the double exponential with density function $\tau \exp(-\tau|\theta|)/2$ and $\delta(0)$ is dirac function with point mass at 0. So if $\gamma_j = 0$, $\theta_j = 0$, and otherwise it is double exponentially distributed with parameter τ. The double exponential is heavier tailed that the normal distribution and therefore can better accommodate large regression coefficients than with the commonly used normal prior $\theta_i|\gamma_i = 1 \sim \mathrm{N}(0, \tau^2)$ (Yuan and Lin 2005). With the double exponential prior, the maximum a posteriori (MAP) estimator is the Lasso estimator (Tibshirani 1996, Park and Casella 2008).

A typical choice of prior for inclusion indicator γ_i is Bernoulli(0.5). Note that in OpenBUGS/ WinBUGS normal distributions are defined in terms of a mean and precision, where precision = 1/variance. The complete BUGS model is given in the Appendix.

Latent environmental variable

So far we considered a single environmental variable denoted by x_i. Community data are multivariate and several environmental factors affect communities (Gauch 1982). There are two ways to extend our model to multiple environmental variables. The first is to extend the quadratic form in Eq. 1 to a general quadratic form, $(\mathbf{x}-\mathbf{u})^t\mathbf{A}(\mathbf{x}-\mathbf{u})$ where **x** and **u** are now vectors with

dimensions associated to the different environmental variables (ter Braak 1988). The second is to stay with Eq. 1 but to redefine x_i as a linear combination of environmental variables, where then the challenge is to find the best linear combination given the data. The first approach uses far more parameters than the second and is more difficult to fit, and for those reasons we use the second approach in this paper. We extend this approach to finding the best sparse combination by applying same Bayesian variable selection approach to the environmental variables as we do for traits. The best sparse linear combination of (measured) environmental can be interpreted as a latent variable driving the phytoplankton communities.

Initial values

We must supply starting values in order to estimate the parameters of a non-linear hierarchical model. Choosing appropriate values can be something of an art. OpenBUGS can crash when inappropriate values are specified.

For obtaining initial values for the Gaussian parameters *a*, *opt* and *tol* for a particular species consider the Gaussian logistic model, that is Eq. 1 without the random site effect,

$$\text{logit}(p) = a - \frac{(x-opt)^2}{2tol^2}, \quad (8)$$

where we dropped the indices for sites and species for convenience. Instead of directly fitting this model to data of a particular species, we rewrite the model as the generalized linear model (ter Braak and Looman 1986, Oksanen et al. 1988) defined as a second-degree polynomial with logarithmic link function

$$\text{logit}(p) = b_0 + b_1 x + b_2 x^2. \quad (9)$$

This model can be easily fitted as a generalized linear model (GLM) with logit link function and, if (estimated) $b_2 < 0$, maximum likelihood estimates of the Gaussian parameters can be found by the following simple formulae (ter Braak and Looman 1986, Oksanen et al. 1988):

$$opt = -\frac{b_1}{2b_2}, tol = \sqrt{-\frac{1}{2b_2}} \text{ and } a = b_0 - \frac{b_1^2}{4b_2}. \quad (10)$$

The coefficients b_0, b_1, and b_2 are thus easily transformed into coefficients representing the species' optimum, tolerance and maximum probability value. The point estimates of the Gaussian parameters thus obtained are identical to those obtained directly using nonlinear maximum-likelihood regression for the Gaussian function. So GLM can be used to derive optimum and tolerance and probability of occurrence that will serve as starting values if $b_2 < 0$. What to do when b_2 is estimated as zero or positive? A standard way is to set b_2 zero and the response curve is in fact sigmoidal. We, instead, simply prevented any nonnegative b_2 by augmenting the data with many zeros (absences) outside the observed range of the environmental variable (at both sides).

We thus viewed such cases as truncated unimodal curves, curves that would have been unimodal if the environmental range in the data were larger. Note the optimum cannot be estimated well if it lies outside or near the edge of the environmental range. By augmenting the data with absences outside the environmental range, the optimum is well-defined and lies within the newly created environmental range. The Bayesian data analysis was, of course, performed on the not-augmented data.

To estimate the initial values for (α_0^a, β_k^a), (α_0^{opt}, β_k^{opt}) and (α_0^{tol}, β_k^{tol}) we regressed the traits on a, opt and tol.

DIC for Model Selection

For comparison of model quality, we use the Deviance Information Criterion (DIC; Spiegelhalter et al. 2002) defined as

$$\mathrm{DIC} = \mathrm{D}(\bar{\theta}) + 2p_D \quad (5)$$

where $\mathrm{D}(\bar{\theta})$ is the posterior deviance evaluated at the posterior mean of the parameter values and p_D the estimated effective number of parameters in the posterior distribution. Spiegelhalter et al. (2002) and OpenBUGS define p_D as the posterior mean of the deviance minus posterior deviance the evaluated at the posterior mean of the parameter values,

$$p_D = \bar{D} - \mathrm{D}(\bar{\theta}) \quad (6)$$

so that

$$\mathrm{DIC} = \bar{D} + p_D \quad (7)$$

Sturtz et al use this equation and approximate p_D as half the posterior variance of the deviance, $p_D = var(deviance)/2$, and estimate it half the average of the within chain variances of the deviance. We used this method for calculating DIC, as it is provided by the R2OpenBUGS function (Sturtz et al.). Eq. (5)1 shows that DIC can be viewed as the Bayesian counterpart to the AIC for model selection. DIC is typically considered as a Bayesian measure of fit or adequacy. The smaller the DIC value, the better the model is.

The DIC statistic is in its early stages and is controversial (Spiegelhalter et al. 2002, Celeux et al. 2006, Gimenez et al. 2009). Here we consider the DIC as a preliminary tool for comparing competing models. As with other model selection criteria, we caution that DIC is not intended for the identification of the best model, but rather merely indicates if a superior model exist within the given candidate models (Huang et al. 2011)

Data Analysis

Data is of 237 species from 211 lakes located within four climate zones (polar to nonpolar) in South America, Europe and North America, and covering a wide range of environmental characteristics. The environmental variables and traits variables are listed in Table 1, which also shows abbreviated names, the unit of measurement, the number of missing values and whether the variable was transformed to natural logarithms in the analysis.

We fitted response models for species which occurred on more than 5% of the sites. The data set is of 203 sites and 60 species. We analysed the data as presence/absence.

Table 1. List of environmental variables and trait variables with code and unit of measurement, number of missing values and indicator for the transformation to natural logarithms.

Variables	Code	Unit	Missing values	Log-transformation
Environmental				
Temperature	Temp	^0C	17	no
Inorganic suspended solids	ISS	$mg\ L^{-1}$	16	yes
Water column mix depth	Zmix	m	2	yes
Light attenuation coefficient	Kd	m^{-1}	4	yes
Conductivity	Cond	µS cm^{-1}	3	yes
Alkalinity	Alk	µeq L^{-1}	8	yes
Total nitrogen	TN	µg L^{-1}	8	yes
Total phosphorus	TP	µg L^{-1}	3	yes
Total zooplankton abundance	TZ	org L^{-1}	8	yes
Cladocera abundance	CLA	org L^{-1}	10	yes
Chlorophyll-a	Chloa	µg L^{-1}	13	yes
Traits				
Volume	V	μm^3	5	yes
Surface area	SV	μm^{-1}	5	yes
Maximum linear dimension	MLD	μm	5	yes
Flagella (presence/absence)	Fla		0	no
Mucilage (presence/absence)	Muc		0	no
Siliceous exoskeleton (presence/absence)	Si		0	no

How to deal with missing values in the trait and environment data? Removing rows (species or sites) with missing values is an option but that means loss of information. Another option is to do imputation. Before imputation, those variables that were clearly not normally distributed were log-transformed to justify the assumption of normality in the imputation procedure (Table 1). Data Imputation was performed using the MICE R-package (Van Buuren and Groothuis-Oudshoorn 2011) using the method "mean" for continuous variables and method "logreg" for binary variables. Finally, each environmental variable and each trait variable was centered and scaled so that the sample mean is zero and the sample standard deviation is 1.

We fitted the Gaussian logistic model to the phytoplankton data with and without trait covariates. We used normal priors and a mixture prior for the regression coefficients. The model was run for each environmental variable for 10,000 MCMC iterations, discarding the first halves as burn-in. For selecting the best sparse linear combination of environmental variables, we ran the Markov chain for 100,000 iterations and discarding the first halves to remove the dependence on the starting values and to allow adequate convergence. In this case convergence of MCMC was very slow.

For all these analysis, the MCMC simulation were performed in the Bayesian software OpenBUGS, linked from the R statistical computing software (R Development Core Team 2010) by R2OpenBUGS (Sturtz et al.). For each analysis, we run three parallel simulation sequences with starting values supplies for some parameters and starting values for other parameters were randomly generated from the prior distributions.

Results

Table 2 shows the model quality in terms of DIC for individual environmental variables and the best linear combination of them (the latent variable). In the both models with and without traits, the latent variable is best, yielding the lowest DIC, followed by chlorophyll-a (Chloa) and temperature (Temp) (Table 2). ΔDIC is the difference between the DIC of a model and the DIC for minimum DIC model. Results shows thus the better predictive ability of chlorophyll-a over that of Temperature. In terms of standardized variables, the latent variable is defined as $(\mathrm{Chloa} - 0.31 \times \mathrm{Temp} - 0.15 \times \mathrm{Zmix} - 0.25 \times \mathrm{Kd} - 0.02 \times \mathrm{Cond} + 0.05 \times \mathrm{Alk} + 0.01 \times \mathrm{TN} + 0.18 \times \mathrm{TZ})$.
From the coefficients of the latent variable model, it appears that the environmental covariates Chloa, Temp, Zmix, Kd and TZ are important. Coefficients for ISS, TP and CLA are zero. Response curves for species along the temperature gradient, Log_{10}Chloa and along the latent variable are plotted in Fig. 1. The species are arranged in ascending order magnitude of their optimum. Species used in analysis along with parameters (a, opt, tol) values obtained from BUGS output for Temperature and Chlorophyll-a are given in Table S1.

Table 2. DIC for individual environmental variables and the best linear combination of them in models with and without trait. The superscripts rank of DIC in ascending order. ΔDIC is the difference between the DIC of model with minimum DIC model.

Env. variable	DIC (without trait)	DIC (trait)	ΔDIC
Temp	6842	6845^{2}	561
ISS	7736	7731^{6}	1447
Zmix	8553	8552^{11}	2268
Kd	8011	8011^{8}	1727
Cond	*	7133^{3}	849
Alk	$6671^{!}$	7244^{4}	960
TN	7816	7891^{7}	1607
TP	8418	8419^{9}	2135
TZ	7442	7442^{5}	1158
CLA	$7738^{!}$	8531^{10}	2247
Chloa	6615	6609^{1}	325
Latent	6283	6284^{0}	0

! negative p_D value; * No convergence

Table 3. Variance components in models without (Null) and with traits (Y&L and normal, indicating the type of prior for trait coefficients). Y&L = Yuan and Lin.

	Variance component	Null	Y&L	Fraction of variation	Normal	Fraction of variation
Temp	σ_a^2	2.10	2.07	1.37	1.93	8.00
	σ_{opt}^2	2.31	1.14	50.45	1.10	52.27
	σ_{tol}^2	0.45	0.15	66.12	0.18	60.80
	DIC	6842	6845		6837.0	
Chloa	σ_a^2	1.56	1.56	0	1.61	-3.39
	σ_{opt}^2	1.77	1.25	29.09	1.32	25.28
	σ_{tol}^2	0.02	0.02	0	0.03	-28.00
	DIC	6615	6609		6596.0	
Latent*	σ_a^2	1.61	1.44	10.72		
	σ_{opt}^2	2.72	2.04	24.89		
	σ_{tol}^2	0.04	0.04	0.00		
	DIC	6283	6284			

* Latent variable, defined as the linear combination of standardized environmental variables
$\text{Chloa} - 0.31 \times \text{Temp} - 0.15 \times \text{Zmix} - 0.25 \times \text{Kd} - 0.02 \times \text{Cond} + 0.05 \times \text{Alk} + 0.01 \times \text{TN} + 0.18 \times \text{TZ}$.

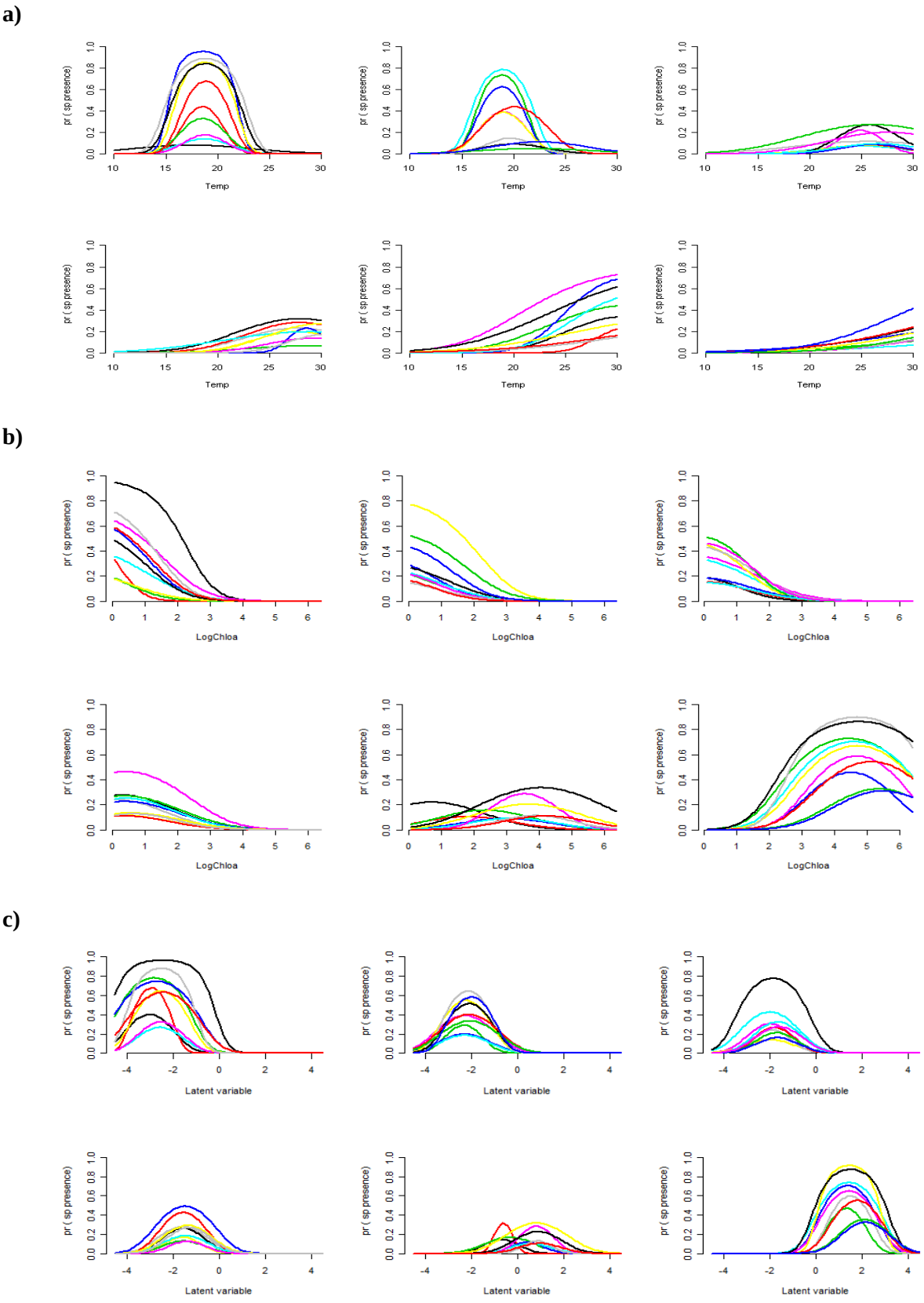

Fig. 1. Response curves for species along the **a)** temperature gradient, **b)** Log(chlorophyll-a) and **c)** the latent variable. Species are arranged in ascending order of their optima values. Each plot has 10 species.

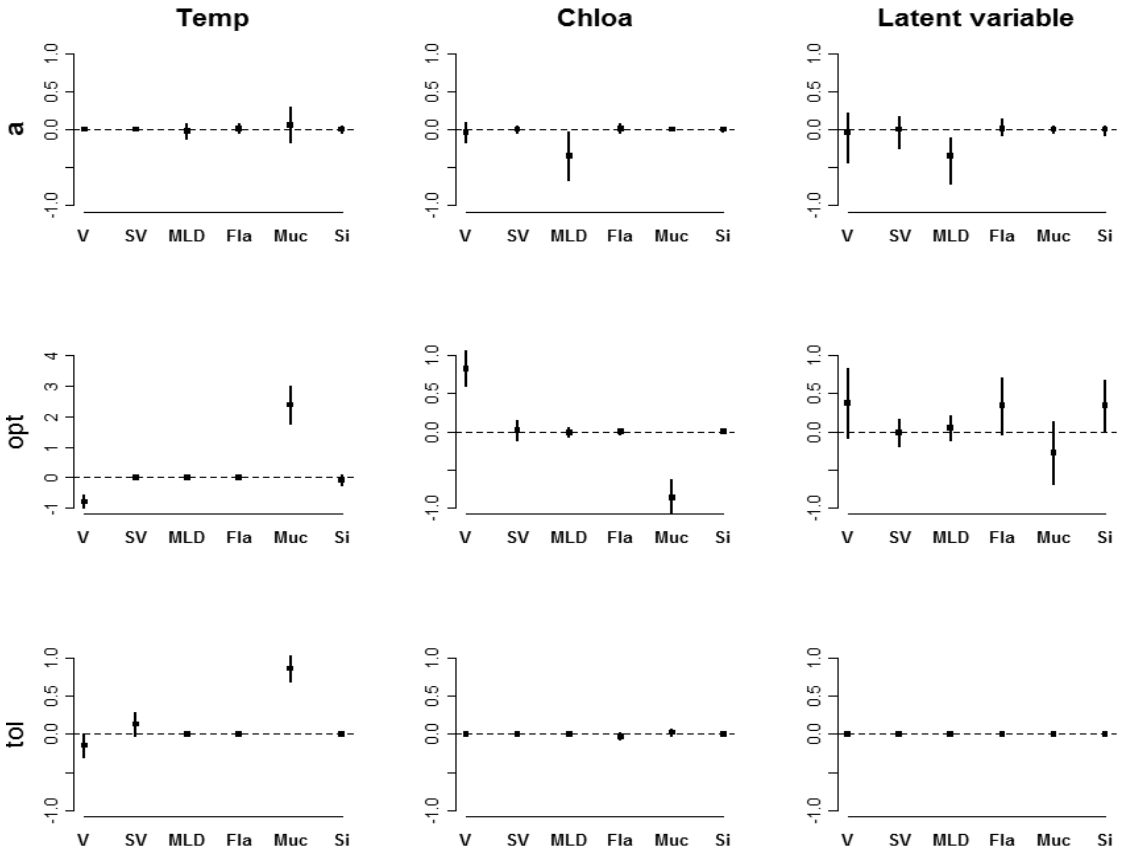

Fig. 2. Coefficient estimate $\pm$ standard deviation for the traits when regressed on Gaussian response parameters (a, opt, tol) for temperature, chlorophyll-a and linear combination of environmental variable.

Table 3 shows percentage variance of the parameters (a, opt, tol) that is explained by the traits using Eq. 5. These parameters were estimated for the best three models where the environmental variable is Temp, Chloa or the latent variable. We observed the percentage variance explained decreases in this order. In our data, the better the environmental variable, the less percentage variance in the parameters is explained by traits.

The traits V, Fla, Muc and Si are important for explaining the variation in optimum. Fig 2 plots the values of the regression coefficients with their standard deviations for the best three models. All trait coefficients for tolerance are zero for Chloa and Latent variable (Fig. 2). It is also evident from Table 3, that traits explain no variation in tolerance for Chloa and the latent variable.

Discussion

Many biotic and abiotic processes contribute to variability in phytoplankton diversity in aquatic ecosystems. The best model constructed in this article reflected well the main mechanisms modulating phytoplankton species growth including temperature (-), resources (light and nutrients: +Chloa, +Alk, -Kd) as well as loss processes (hydrological washout, sedimentation and consumption by zooplankton: -Zmix, +TZ) (Margalef, 1978; Reynolds, 1984).

Temperature has important direct effects on phytoplankton growing rates (Reynolds 1984a, Tilman 1986) and indirect effects changing water properties and water column mixing. Temperature affects latitudinal distribution of species (Fuhrman et al. 2008) and the organisms metabolism (Brown et al. 2004).

Of all variables included in our study, chlorophyll-a had the largest effect. It is a measure of total phytoplankton biomass and not an environmental variable *per se*. However, it reflects a combination of variables related to the trophic state of the lakes, and therefore recourses (nutrients, light) and ecosystem productivity. Higher chlorophyll-a is usually related to high nitrogen and phosphorus in the silicate and carbon, and to the effect of the watershed, increased alkalinity is associated to higher nutrients concentrations (Conley 2002).

Depth of the mixing zone is an important indicator of the phytoplankton environment particularly in relation with light availability and sedimentation losses. High mixing depth produce potentially higher times for the phytoplankton in depths of the water column with low light and is related to mixing losses (Reynolds 1997).

Finally, total zooplankton abundance (TZ) includes the three dominant groups of zooplankton in lakes pelagic zones: rotifers, copepods and Cladocera. Cladocera are the main phytoplankton controllers but are dominant under oligo to mesotrophic conditions. Rotifers are also good filter-feeders but are dominant in more eutrophic conditions attaining the highest abundance therefore affecting drastically this community total abundance (Reynolds 1984b, Lampert and Sommer 2007).

Interestingly, the two best models with one variable representing a temperature and productivity/trophic state gradient are the focus of intensive research nowadays. Chlorophyll-a and temperature are used as indicators of eutrophication and climate warming. Those processes affect dramatically aquatic ecosystems nowadays modifying their communities and functioning, promoting the species invasion and modifying the trophic interactions (Thuiller 2007, Paerl and Huisman 2008, Paerl and Huisman 2009, Kosten et al. 2011, Moss et al. 2011).

An interest result of our analysis is that differently from expectations (Moss et al. 2011) temperature and productivity gradients showed an opposite effect. The explanation might be linked to differences in the relative importance of temperature and nutrients along latitudinal gradients (Kosten et al. 2009a, Kosten et al. 2009b). Also it might be caused by differences in the effects of trophic interactions between warmer and cooler lakes (Malthus and Mitchell 2006, Kosten et al. 2009b).

How different traits (V, Fla, Muc, SV, Si) affect the species habitat features (optimum, tolerance and maximum)?

Phytoplankton morphological traits reflect the ability to acquire resources (light and nutrients), to grow and to avoid mortality, through such processes as hydrological washout, sedimentation and consumption by grazers (Margalef 1978, Reynolds 1984b). The relation between morphological traits and physiology is well-defined for phytoplankton (Lewis 1976, Margalef 1978, Reynolds 1988, Elliott et al. 2001, Kruk et al. 2010).

Volume and surface/volume ratio affect specific growth rate, resource-uptake and light-interception properties (Reynolds 1988, Kirk 1996, Kruk et al. 2010). In general terms smaller size and higher SV potentiate higher growth rates and a greater tolerance to limiting light conditions (Naselli-Flores and Barone 2007). Therefore we would expect organisms with smaller volume to attain their optimum distribution at lower values along the trophic/productivity gradient (chlorophyll-a), as was observed in our results. Larger volumes will be expected also at the end of succession when higher biomass is attained in the community and higher nutrients are available, therefore increasing the optimum values as has been observed in other studies (Sommer 1989, Kruk et al. 2002).

Volume and surface/volume ratio affect specific growth rate, resource-uptake and light-interception properties (Reynolds 1988, Kirk 1996, Kruk et al. 2010). In general terms smaller size and higher SV potentiate higher growth rates and a greater tolerance to limiting light conditions (Naselli-Flores and Barone 2007). Therefore we would expect organisms with smaller volume to attain their optimum distribution at lower values along the trophic/productivity gradient (chlorophyll-a), as was observed in our results. Larger volumes will be expected also at the end of succession when higher biomass is attained in the community and higher nutrients are available, therefore increasing the optimum values as has been observed in other studies (Sommer 1989, Kruk et al. 2002).

The relation with mucilage is not so clear. The presence of mucilage provides controllable buoyant properties (Ferber et al. 2004), may help maintaining an adequate microenvironment for cells and avoidance of grazing Reynolds, 2007 (Reynolds 2007)). Also, survival may be prolonged by the facility of remaining as resting colonies in the sediment (Reynolds 1981). The inverse relation between volume and mucilage in relation with chlorophyll-a might be the following. Biovolume is a measure of phytoplankton biomass comparable to chlorophyll-a is and is calculated as the volume of individual organisms, estimated from approximated geometrical shape, multiplied by their abundance in the environment (mm^3L^{-1}). Therefore higher concentrations of chlorophyll a are related to higher biovolume of the phytoplankton community. However, mucilage does not

contribute to biovolume in terms of photosynthetically active biomass, therefore increasing volume in terms of mucilage (e.g. sorrowing cells in a colony) might be related to lower chlorophyll.

The case of the latent variable represents a gradient from lower to higher chlorophyll along with higher to lower temperature, Kd, Zmix and total zooplankton abundance. Volume increased the optimum and the same reasoning applies as before in relation with chlorophyll-a. Larger volumes increase the optimum of the species along the productivity gradient. In the case of the latent variable also the presence of flagella increased the optimum, this might be explained because of the motility might allow algae to forage for nutrients and avoid grazing (Reynolds 1997). Increasing productivity gradient was accompanied by high total zooplankton abundance which is also typical of more eutrophic systems, which are usually dominated by smaller zooplankton species organisms like rotifer (Lampert and Sommers 2007). Furthermore, grazing efficiency by filter-feeding zooplankton is affected by phytoplankton size and morphology, being larger organisms less edible (Burns 1968, Lampert 1987, Lehman 1988, Reynolds 2006). Smaller organisms survive high grazing pressure due to their higher growth rates, and therefore smaller volumes would decrease the species optimum along the latent variable gradient. Higher Kd interpreted as lower light in the water column, and lower mixing depth would favours a lower time under light limitation conditions. Limitation by light would force the organism to increase their S/V and therefore their light reception capabilities, which is larger in organisms with lower volume (Lewis 1976). Size also affect sinking losses, and species responses to disturbance (Reynolds 1984b, Padisák 2003). Smaller and high S/V organisms sink slowly, and survive high water flushing.

A general different effect of the traits in the allocation of the optimum distribution was observed for temperature, as was also observed for the latent variable. Direct effect of temperature in organisms includes the acceleration of their metabolism, increasing their growth rates (higher C assimilation), their senescence rate (higher photo-respiration) and therefore decreasing their average size. Therefore at higher temperatures we would expect smaller size and volume, along with higher S/V. The negative effect of temperature in size was also observed in paleoecological studies (Smol et al. 2005; Ruhland et al. 2008) and actual field analysis (Winder et al. 2009). In those cases the effect was mainly indirect through the effect of higher temperatures in water properties and aquatic ecosystems mixing regime. Smaller organisms with higher S/V sink slowly therefore the presence of smaller organisms is favoured (Winder et al. 2009). Finally, the presence of siliceous structures affected the location of the optimum increasing its position along the latent variable. The obligate presence of a siliceous wall affect cell density and organisms sink rapidly and are excluded from waters depleted in assimilable sources of silica (Padisák 2003). Furthermore, siliceous walls also have advantages against certain types of grazers (Hamm et al.

2003) and viral infections (Smetacek 2001) and the presence of siliceous spines might reduce losses because of grazing (Reynolds and Irish 1997).

The use of functional traits improved the performance of the models only including habitat characteristics. In our data we made the surprising observation that the better the environmental variable was in explaining the habitat template, the less variation in the parameters was explained by traits.

Another interesting aspect is that we only included environmental variables associated to local environments in our analysis. The habitat template should also include variables associated to species distribution and regional or global processes. Here the inclusion of functional traits as volume and shape is directly related to distribution processes and including it in the habitat template might correct for this limitation (Fenchel and Finlay 2004).

Groups in the response curves along the habitat templates

A final striking result is that we observed two groups of species along the three habitat template models (Fig. 3). This might recall the idea of communities as functional entities (Clements 1916) and the phytosociological approaches (Braun-Blanquet 1964), as well as the classical question of how many species can co-exist (Hutchinson 1961). A recent alternative explanation for the co-existence of many species is advocated by the combination of neutral theory of biodiversity (Hubbell 2001) and niche theory. The theory of self-organized similarity (also referred to as 'Emergent neutrality') proposes that there may be a limited number of evolutionary self-organized functional groups of species (and corresponding niches), but that within each group an essentially unlimited number of ecologically equivalent species might co-exist neutrally (Scheffer and van Nes 2006). Nowadays new studies are recognizing this theory as potential explanation (Vergnon et al. 2009, Segura et al. 2011) but still more studies are needed.

Conclusion

This paper presents a Bayesian approach for modelling a unimodal species response model relating traits to environment from phytoplankton communities. Species response curves showed that species are divided into clusters and the variation within cluster seems very low. DIC was useful to select the potentially important environmental variables, but less useful to select potentially important traits because no important difference in DIC was observed between the models with and without traits. The variation in the niche parameters (a, opt, tol) explained by species traits was measured by the contribution statistic C_β. About 25% of the variance in the niche centres with respect to chlorophyll-a could be accounted for by the traits, whereas niche width and maximum could not be predicted. Volume, mucilage and flagella are found to be the most important traits to explain the niche differences.

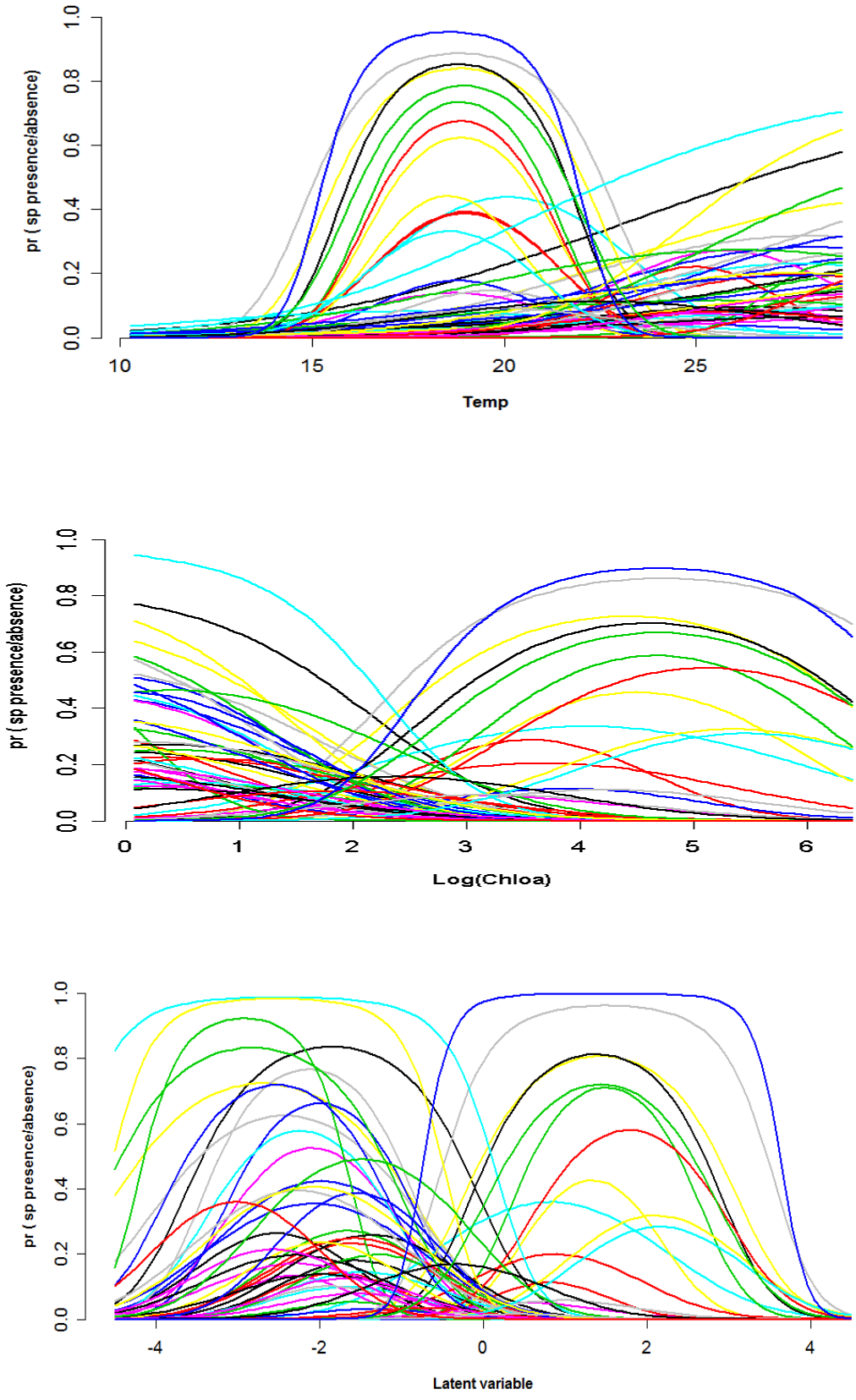

Fig. 3. Response curves for species along the temperature gradient, Log(chlorophyll-a) and the latent variable.

Of course, not all measurable features are equally important and some important features may perhaps be combined into a synthetic (latent) environmental gradient. It is formed by a linear combination of environmental variables that are presumed to maximally explain the species distribution.

We fitted the models in a fully Bayesian approach, employing the MCMC simulation to generate posterior samples from the joint posterior distribution, which can be used to make various posterior inferences. Although Bayesian methods are computationally intensive, they are easy to implement and provide not only point estimates but also interval estimates of all parameters. The fully Bayesian approach enabled us to obtain much richer inferences about the models than most non-Bayesian analyses.

We assumed that species response on an environmental gradient has a symmetrical bell shaped (Gaussian) curve. However other types of response also occur quite common because interactions between species and extreme environmental stress may cause skewed or non-unimodal responses. The Bayesian approach can be extended to other parametric nonlinear models with parameters made dependent on traits.

Appendix

Bugs model for Latent variable (with traits)

```
#############################################################################
###
# X0=Chloa: X1=Temp; X2=ISS; X3=Zmix; X4=Kd; X5=Cond; X6=Alk;, X7=TN; X8=TP; X9=TZ; X10=CLA
# Z= (V, SV, MLD, Fla, Muc, Si)
# opt=Optimum; tol= Tolerance; logc= a
#############################################################################
```

```
 model{                        # N observations (species*sites)
   for (i in 1:N){

      y[i] ~ dbin(p.bound[i],1)
      p.bound[i] <-max(0,min(1,p[i]))
      logit(p[i])<-Xbeta[i]
      Xbeta[i]   <-logc[sp[i]]-0.5*pow((Xstar[i]-
opt[sp[i]])/tol[sp[i]],2)+
                    b.site[site[i]]
      Xstar[i]<-beta0*X0[i]+beta[1]*X1[i]+beta[2]*X2[i]+beta[3]*X3[i]+
                beta[4]*X4[i]+beta[5]*X5[i]+beta[6]*X6[i]+beta[7]*X7[i]+
                beta[8]*X8[i]+beta[9]*X9[i]+ beta[10]*X10[i]
      }
```

```
for (j in 1:n.sp)  {

       logc[j] ~ dnorm(a.hat[j],tau.a)
       a.hat[j]<-a0+inprod(a[],Z[j,])

       opt[j] ~ dnorm(opt.hat[j],tau.opt)
       opt.hat[j]<-b0+ inprod(b[],Z[j,])

       tol[j] ~ dnorm(tol.hat[j],tau.tol)
       tol.hat[j]<-c0+ inprod(c[],Z[j,])

      }

   for (j in 1:n.site) {
       b.site[j] ~ dnorm(0,tau.site)
      }

   for ( k in 1:n.env){
       beta[k]~ddexp (0,taubeta[k])
       taubeta[k]<-1/varbeta[k]
       varbeta[k]<-(1-gammbeta[k])*0.001+gammbeta[k]*10
       gammbeta[k]~dbern(pi.beta) }

    beta0<-1
    a0 ~dnorm(0,.0001)
    b0 ~dnorm(0,.0001)
    c0 ~dnorm(0,.0001)

   for ( k in 1:6){
       a[k]~ddexp (0,taua[k])
       taua[k]<-1/vara[k]
       vara[k]<-(1-gamma[k])*0.001+gamma[k]*10
       gamma[k]~dbern(pi.a)

       b[k]~ddexp (0,taub[k])
       taub[k]<-1/varb[k]
       varb[k]<-(1-gammb[k])*0.001+gammb[k]*10
       gammb[k]~dbern(pi.b)

       c[k]~ddexp (0,tauc[k])
       tauc[k]<-1/varc[k]
       varc[k]<-(1-gammc[k])*0.001+gammc[k]*10
       gammc[k]~dbern(pi.c)
              }

     tau.a<-pow(sigma.a,-2)
     sigma.a~dunif(0,100)
     tau.opt<-pow(sigma.opt,-2)
     sigma.opt~dunif(0,100)
     tau.tol<-pow(sigma.tol,-2)
     sigma.tol~dunif(0,100)
     tau.site<-pow(sigma.site,-2)
     sigma.site~dunif(0,100)

     pi.a<-0.5
     pi.b<-0.5
     pi.c<-0.5
     pi.beta<-0.5
   }
```

Table S1. Species names and parameters (a, opt, tol) values obtained from BUGS output for Temperature and Chlorophyll-a

		Temperature			Chlorophyll-a		
	Species name	**A**	**Opt**	**Tol**	**A**	**Opt**	**Tol**
1	Aphanocapsadelicatissima	1.19	42.09	10.05	1.36	-0.74	1.50
2	Aphanocapsaholsatica	-1.19	40.82	9.91	-0.66	-0.86	1.32
3	Aphanocapsaincerta	-0.21	46.39	11.51	-0.67	-0.40	1.46
4	Aphanotheceminutissima	-0.68	37.76	9.70	-0.24	-0.33	1.38
5	Aulacoseiragranulatavgranulata	-0.25	20.10	2.45	-0.68	4.06	1.59
6	Chlorellahomosphaera	-0.99	25.72	2.58	-0.21	-0.44	1.30
7	Chlorellaminutissima	-0.22	31.62	6.33	0.84	-1.05	1.52
8	Chlorellavulgaris	-0.77	27.92	5.25	0.16	-0.50	1.50
9	Chlorococcales4	-2.24	26.62	3.83	-1.11	-0.74	1.36
10	Chroomonassp	-0.46	18.99	1.90	-0.91	3.55	0.91
11	Chrysococcussp	1.02	18.85	1.46	0.35	4.68	1.03
12	Cryptomonasbrasiliensis	-0.93	27.94	4.59	0.06	-0.15	1.10
13	Cryptomonasmarsoniipeq	-1.21	27.70	4.66	-0.16	-0.34	1.19
14	Cryptomonasmarssonii	-2.35	26.21	3.04	-1.44	-0.20	1.12
15	Cryptomonassp	1.66	18.84	1.75	0.98	4.42	1.19
16	Cyanodictyonimperfectum	1.17	47.66	10.16	0.67	-1.07	1.31
17	Cyclotella/Stephanodiscus	1.76	18.79	1.50	0.86	4.60	1.17
18	Cyclotellamengehiniana	-1.26	24.93	1.93	-1.23	0.37	1.49
19	Cylindrospermopsisraciborskii	-1.58	30.30	3.16	-1.93	0.64	1.25
20	Dinobryondivergens	-2.95	22.08	5.37	-1.50	-0.51	1.15
21	Dyctiosphaeriumpulchellum	-0.76	45.39	9.86	-1.75	-0.11	1.46
22	Epithemiasp	-1.82	18.69	2.36	-2.32	3.00	1.32
23	Euglenasp	0.51	18.87	1.58	-0.17	4.49	1.05
24	Eukaryioticnanoplankton	2.06	18.81	1.83	1.82	4.73	1.19
25	Eutetramorusfotii	-1.02	42.45	11.14	-1.14	0.04	1.40
26	Gomphonemasp	-0.44	18.98	1.89	-1.36	3.67	1.48
27	Gymnodiniumcnecoides	-1.19	28.60	1.57	0.53	-1.36	0.91
28	Gyrosigmasp	-1.52	18.78	1.88	-2.05	3.89	1.10
29	Jaaginemagracile	-2.42	17.51	5.48	-1.86	0.56	1.43
30	Lemmermmaniellapallida	-0.94	43.70	9.29	-1.96	-0.02	1.57
31	Lepocinclissalina	-0.23	18.54	1.54	-0.73	5.35	1.27
32	Merismopediaduplex	-0.30	47.47	10.19	-1.09	-0.89	1.44
33	Merismopediatenuissima	-0.08	46.81	11.45	-1.00	0.35	1.52
34	Monoraphidiumcontornum	-1.41	27.67	3.88	-1.27	0.71	1.45
35	Monoraphidiumconvolutum	-2.03	25.55	6.61	-1.09	0.38	1.38
36	Oocystislacustris	-0.87	43.42	12.25	-0.32	-1.06	1.58
37	Oocystismarsonii	-1.58	42.79	12.29	-1.08	-0.79	1.51
38	Oocystisparva	-1.91	43.59	11.85	-1.62	-0.72	1.45
39	Oocystissp.1	-0.45	45.22	10.62	-0.93	-0.52	1.47
40	Peridiniumsp	-1.76	19.64	2.14	-2.09	4.22	1.26
41	Peridiniumumbonatumvumbonatum	-2.38	25.86	3.14	-1.67	-0.19	1.37

42	Phacussp	0.74	18.84	1.52	0.17	5.11	1.23
43	PicoChloroesférico	0.23	33.16	5.14	0.63	-0.93	1.33
44	PicoChloroesferóide2	-0.69	30.61	4.66	0.23	-0.95	1.33
45	Picocyano(<1um)	1.17	34.64	7.60	3.40	-1.40	1.36
46	PicoCyanocilindrico2	-2.52	24.88	3.93	-1.24	-1.03	1.43
47	PicoCyanoesférico1	0.92	32.10	4.22	1.25	-1.01	1.28
48	Planktolyngbyalimnetica	-0.99	29.97	4.46	-0.95	0.22	1.45
49	Pseudanabenarecta	-2.68	28.47	5.49	-2.07	0.31	1.35
50	Raphidiopsismediterranea	-2.43	25.49	3.80	-2.14	1.81	1.32
51	Rhodomonasminuta	-0.98	26.14	6.09	-0.15	0.41	1.44
52	Scenedesmusellipticus	-1.41	28.95	7.64	-0.16	-0.07	1.19
53	Strombomonassp	-0.70	18.56	1.88	-0.80	5.46	1.26
54	Synechococcusaquatilis	-1.86	29.11	4.12	-1.48	-0.12	1.41
55	Synechococcusnidulans	-1.38	27.53	5.22	-0.56	-0.39	1.58
56	Synedraacus	-2.30	19.97	3.03	-2.29	3.20	1.27
57	Tetraedronminimum	-2.06	22.51	4.23	-1.68	2.36	1.36
58	Tetraëdroncaudatum	-1.22	30.84	2.60	-0.97	-1.25	1.31
59	Trachelomonassp	1.30	18.94	1.65	0.70	4.68	1.17
60	Trachelomonasvolvocina	3.01	18.58	1.34	2.16	4.69	0.97

Chapter 4

A Generalized Linear Mixed Model approach to species-environment relationships can handle and detect unimodal relationships with simulated and real data examples

Tahira Jamil, Cajo J.F. ter Braak

Abstract

Niche theory predicts that species occurrence and abundance show non-linear, unimodal relationships with respect to environmental gradients. Unimodal models, such as the Gaussian (logistic) model, are however much more difficult to fit to data than linear ones, particularly when also species phylogeny and species traits are to be taken into account. This is one of the reason for the popularity of canonical correspondence analysis and RLQ in ecology. These methods are very useful with unimodal data but are linear after transformation. This paper explains why and when generalized linear mixed models can effectively analyse unimodal data and also presents a graphical tool and statistical test to test for unimodality while fitting just a generalized linear mixed model.

Key-words: Niche theory; environmental gradient; testing unimodality; Gaussian logistic mode; Generalized linear mixed model

Introduction

Niche theory predicts that species occurrence and abundance show non-linear, unimodal relationships with respect to environmental gradients (Økland 1986, Austin 1987, Minchin 1989, Palmer and Dixon 1990). Many studies fail to test for unimodal response (Austin 2007). Thus straight-line relationships are often fitted without justification (e.g. Gibson et al. (2004)).

Ordination is a class of multivariate methods to analyze the occurrence and/or abundance of a set of species in a set of sites and results in a configuration of the sites in a factorial plane, the directions of which can be interpreted as latent environmental variables (Jongman et al. 1995, ter Braak and Prentice 2004, Walker and Jackson 2011). Principal component analysis, (detrended) correspondence analysis are rival eigen vector methods for this. In constrained or canonical ordination, the latent variables may be constrained to linear combination of manifest (measured) environmental variables, and the above rival methods become redundancy analysis and canonical correspondence analysis (ter Braak and Verdonschot 1995), together with alternatives such as coinertia analysis (Dolédec and Chessel 1994). With species traits in the analysis the latter becomes the RLQ method (Dolédec et al. 1996, Bernhardt-Romermann et al. 2008). Principal component analysis and redundancy analysis are known as the linear methods whereas correspondence analysis, and canonical correspondence analysis are claimed to be an approximation to fully unimodal methods (ter Braak 1987). Nevertheless, (canonical) correspondence analysis is an eigen vector method and therefore inherently linear. This is most apparent in the reconstitution formula of (canonical) correspondence analysis (Greenacre 1984, ter Braak and Verdonschot 1995). How can it be understood that these methods are able to model unimodal data but are inherently linear? The same question can be phrased for principal components analysis on transformed data, such as double centered, log transformed data, or data standardized to equal site total or equal site norm.

Some insight in this question is given by Ihm and van Groenewoud (1984) and further worked out by ter Braak (1987) and de Rooij (2007) who show the relation between the unimodal model and Goodman's RC model, which is a generalized linear model, and a loglinear model in particular. The relation can be used both ways. Ihm and van Groenewoud (1984) use the relationship to justify the RC model (there called Model B) for ecological ordination and de Rooij (2007) uses it to transform the linear predictor of the RC model into a quadratic form, with the graphical purpose to transform a vector representation or biplot to a distance representation that is supposed to be easier to interpret for naïve users of multivariate methods.

In this paper we use the same approach to derive a graphical tool and statistical test to test for unimodality while fitting just a generalized linear mixed model (GLMM). GLMMs are model-

based, inferential statistical tools for describing the underlying community pattern and are becoming popular in ecological and evolutionary studies (Bolker et al. 2009, Ives and Helmus 2011). We claim that GLMMs can effectively analyze unimodal data when the niche width is not very different among species and illustrate this claim by comparing the GLMM approach with an explicit unimodal model approach on data that show unimodality.

Theory

Unimodal curves and generalized linear (mixed) models

For easy of exposition we use logistic linear (mixed) models as example of generalized linear (mixed) models. The same approach can be followed for loglinear model, which would relate to the RC model (de Rooij 2007). One of the simplest unimodal curves for presence-absence data is the Gaussian logistic curve (ter Braak and Looman 1986)

$$\text{logit}(p_{ij}) = a_j - \frac{(x_i - u_j)^2}{2t_j^2} \qquad (1)$$

with p_{ij} is the probability of occurrence [p_{ij} = E(y_{ij}), the expected value of the observation y_{ij}], x_i a quantitative known environmental variable, a_j is a coefficient related to maximum probability of occurrence, u_j is the species optimum and t_j is the tolerance of species *j*. The subscripts *i* and *j* refer to site *i* and species *j* respectively (*i*=1,…, *n*; *j* = 1,…, *m*). This model has thus a logistic form and is nonlinear in this parameterization. By expanding the quadratic term in equation (1) and assuming $t_j = t$, we obtain

$$\begin{aligned} a_j - \frac{(x_i - u_j)^2}{2t^2} &= a_j - \frac{1}{2t^2}x_i^2 - \frac{1}{2t^2}u_j^2 + \frac{1}{t^2}x_i u_j \\ &= \left(a_j - \frac{1}{2t^2}u_j^2\right) + \left(\frac{u_j}{t^2}\right)x_i - \frac{1}{2t^2}x_i^2 \ . \end{aligned} \qquad (2)$$

By setting

$$\alpha_j = a_j - \frac{1}{2t^2}u_j^2,\ \beta_j = \frac{u_j}{t^2} \text{ and } \gamma_i = -\frac{1}{2t^2}x_i^2, \qquad (3)$$

we obtain a Generalized Linear Model (GLM)

$$\text{logit}(p_{ij}) = \alpha_j + \beta_j x_i + \gamma_i \qquad (4)$$

where α_j is an intercept, β_j a slope and γ_i a site effect. If t would vary among species then equation (3) does not exactly hold because x_i^2/t_j^2 then also depends on *j*. We will turn this GLM into a Generalized Linear Mixed Model (GLMM) by assuming that the three parameters are random effects deriving from three distributions, for which we will take normal ones for numerical

convenience. The random site effects are of the form $-x^2$, which has a nonzero mean. This mean can be taken out and transferred to the mean of the distribution of the intercepts $\{\alpha_j\}$, so the distributional assumptions are

$$\binom{\alpha_j}{\beta_j} \sim N\left(\binom{\alpha_0}{\beta_0}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix}\right) \quad (5)$$

where σ_α^2 and σ_β^2 are the variance components for α_j and β_j and ρ is the correlation between α_j and β_j, and

$$\gamma_i \sim N(0,\ \sigma_\gamma^2) \quad (6)$$

with σ_γ^2 the variance component for the site effects.

There is nothing special about equation (1) stating a single environmental variable. The model can be extended to two environmental variables (ter Braak and Prentice 2004)

$$\text{logit}(p_{ij}) = a_j - \frac{1}{2}\Big(d_1(x_{i1} - u_{1j})^2 + d_2(x_{i2} - u_{2j})^2 - 2d_{12j}(x_{i1} - u_{1j})(x_{i2} - u_{2j})\Big). \quad (7)$$

where d's are precision parameters, in the context of the bivariate normal distribution (Rue and Held 2005). By setting

$$\alpha_j = a_j - \frac{1}{2}(d_1u_{1j}^2 + d_2u_{2j}^2 - 2d_{12j}u_{1j}u_{2j}),$$

$$\beta_{1j} = d_1u_{1j} - d_{12j}u_{2j},$$

$$\beta_{2j} = d_2u_{2j} - d_{12j}u_{1j},$$

$$\beta_{3j} = d_{12j}, \quad (8)$$

and

$$\gamma_i = -\frac{1}{2}(d_1x_{i1}^2 + d_2x_{i2}^2) \quad (9)$$

we can write

$$\text{logit}(p_{ij}) = \alpha_j + \beta_{1j}x_{i1} + \beta_{2j}x_{i2} + \beta_{3j}x_{i1}x_{i2} + \gamma_i\ . \quad (10)$$

Here β_{3j} are random effects for interactions. If the 'co-precisions' are equal ($d_{12j} = d_{12}$) the term $\beta_{3j}x_{i1}x_{i2}$ can be subsumed in to the site effects $\{\gamma_i\}$ and the model can do without interactions. The $\{\gamma_i\}$ account for the quadratic term arising from the Gaussian (logit) model. In conclusion, up to distributional assumptions, the GLMM can be interpreted as a Gaussian logit model with equal tolerances for the species.

A graphical tool and statistical test for unimodality

Equations (3) and (9) suggest a graphical tool for detecting unimodality and also a statistical test. The idea is to fit a GLMM to the binary data $\{y_{ij}\}$ with respect to the environmental variable with values $\{x_i\}$ ($i = 1,\ldots, n$). In the R package lme4 (Bates et al. 2011), the model can be fitted by

```
lmer (y ~ 1+ x + (1+ x | sp) + (1| site)
          family=binomial(link="logit") ,data),
```

where y represents the vectorized response data while sp and site are factors indicating species and sites. The site effects $\{\gamma_i\}$ obtained from the fit are then plotted against the environmental variable $\{x_i\}$. There is an indication of unimodality in terms of the species response with respect to the environmental variable x if this graph shows a quadratic relationship. An associated statistical test can be obtained by regressing the site scores on x and x^2, i.e. using linear multiple regression with model formula

$$site.score \sim x + x^2 \qquad (11)$$

and examining the significance of the squared term by a t-test on its regression coefficient.

The usual role of the site effects in a GLMM such as equation (4) is to account for the size of the site or the fertility of the site and, in general, for factors that influence the probability of occurrence of all species in the site. The site effect γ_i will thus be expected to be related to the expected number of species in a site, that is to $\sum_j p_{ij}$ and, in terms to the data to the number of species that is observed in a site, for short the site total, defined as $S_i = \sum_j y_{ij}$. The site total and the site score are thus naturally related. In order to obtain a more sensitive test, it is thus logical to add the site total to the formula in in equation (11), giving

$$site.score \sim x + x^2 + S. \qquad (12)$$

where S is the site total (number of species in a site). There is evidence of unimodality if the squared term is significant as judged by a t-test on its regression coefficient.

Material and methods

Simulation Set-up

Example 1: The procedure to simulate data was the following:

1) Generate n=50 values of an environmental variable x as a random sample from the uniform distribution such that $x \sim U(-2\,, 2)$.
2) Generate a vector $\boldsymbol{a}$, a parameter related to maximum probability of length m (number of species) drawn at random from a normal distribution with 0 mean and unit variance ($a \sim N(0,1)$).

3) Generate a vector u of length *m* from a uniform distribution such that $u \sim U(-\tau, \tau)$, where $\tau = 2 + t$, for a fixed value of *t*, to ensure that optima are also placed outside the sample range of *x*.
4) Generate the binomial probabilities form the unimodal response curve

$$p_{ij} = \text{logit}^{-1}\left(a_j - \frac{(x_i - u_j)^2}{2t_j^2}\right) \quad (13)$$

where y_{ij} presence–absence data were generated at random from a binomial distribution with probability p_{ij} and $t_j = t$. We simulate data for constant tolerance in each data set with *m*=100 species for $t = (0.5, 1, 2, 4)$. Fig. 1 indicates how the simulated species response curves look like for different values of tolerances.

Example 2:

Example 2 is as example 1, except that it uses a normal instead of a uniform distributions for *x* and *u*, $x \sim N(0, 1)$ and $u = u^* \times t$ where $u^* \sim N(0, 1)$.

Example 3:

In this example the tolerance varies among species, with a median tolerance of 0.5 in example 3A and of 1 in example 3B. Let $\sigma = (0.1, 0.25, 0.5, 1)$, for each σ generate a vector t of length *m*=100 from a lognormal distribution $t \sim \text{LogN}(\mu, \sigma)$, where $\exp(\mu) = 0.5$ and 1 in examples 3A and 3B, respectively. The rest of the setup is the same as example 1.

Example 4:

In this example we simulate datasets with fewer (*m*=10) and more (*m*= 100) species than in example 1 (*m*= 50) with tolerance t =1. Simulation setup is the same as example 1.

Each dataset was characterized by beta diversity and length of gradient The most commonly used index of beta diversity is $\beta_w = S/\bar{\alpha} - 1$, where *S* is the total number of species, and $\bar{\alpha}$ is the average number of species per site (Whittaker 1960). Length of gradient is a property of an environmental variable. The length of gradient can be defined as the range of the environmental variable divided by the average range of the species. The axis length of first axis of detrended correspondence analysis (DCA) of data is the Length of gradient (ter Braak 1993). Length of gradient was expressed in standard deviation (S.D.) units (Hill and Gauch 1980). Beta diversity was calculated using the asbio package (Aho 2011) and DCA was performed in the *vegan* package (Jari et al. 2011), both in R software (R Development Core Team R 2011). We also simulated data according to GLMM model of equations (4)-(6).

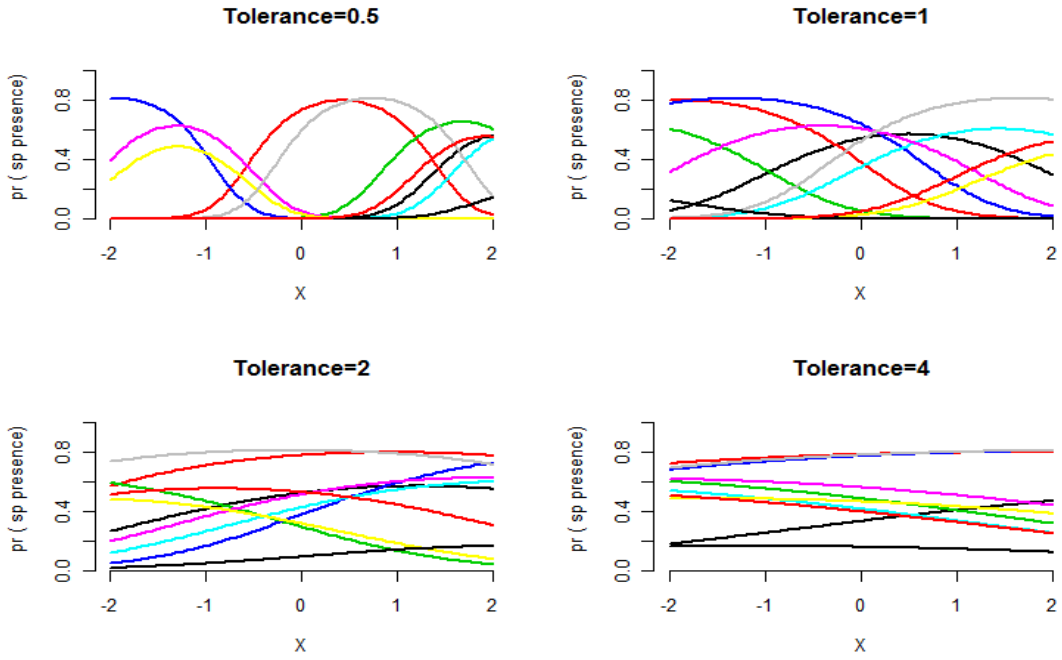

Fig. 1. Simulated unimodal response curves of the probability that species occur at a site, against the environmental variable x for 4 different values of tolerances $(0.5, 1, 2, 4)$. For each tolerance value, curves vary in optimum and maximum probability of occurrence.

Results and discussion

Simulated data

Fig. 1 shows the simulated response curves in examples 1 and 2; the sampled range of the environmental variable is the range of x shown. With increasing tolerance the part of the curves that is sampled shows less unimodality. This is expressed quantitatively in Table 1 by the length of gradient SD units which varies between about 1 SD (not so unimodal) to 6 SD (very unimodal). The associated beta diversity varied between 1 and 5.

The site effects issued by the GLMM analysis of each of the simulated data sets are plotted in Figs. 2-6 against the environmental variable (left) and the site total (middle), together with a plot of the site total against the environmental variable (right).

For example 1 (uniform set-up), site effects shows a clear quadratic relationship along the environmental variable for t = 0.5, 1, and 2 but for $t = 4$ the relationship does not look very quadratic (Fig. 2). The range of sites effects also decreases with increasing tolerance and for large t, the site effects are close to zero, as can be seen from the vertical scale values. The plot of the site effects against the site total is linear but more dispersed for small tolerance than for large tolerance. For large tolerance they become nearly exactly linear (middle column of Fig. 2). The site total has also a weak quadratic relationship with the environmental gradient but for large tolerance no

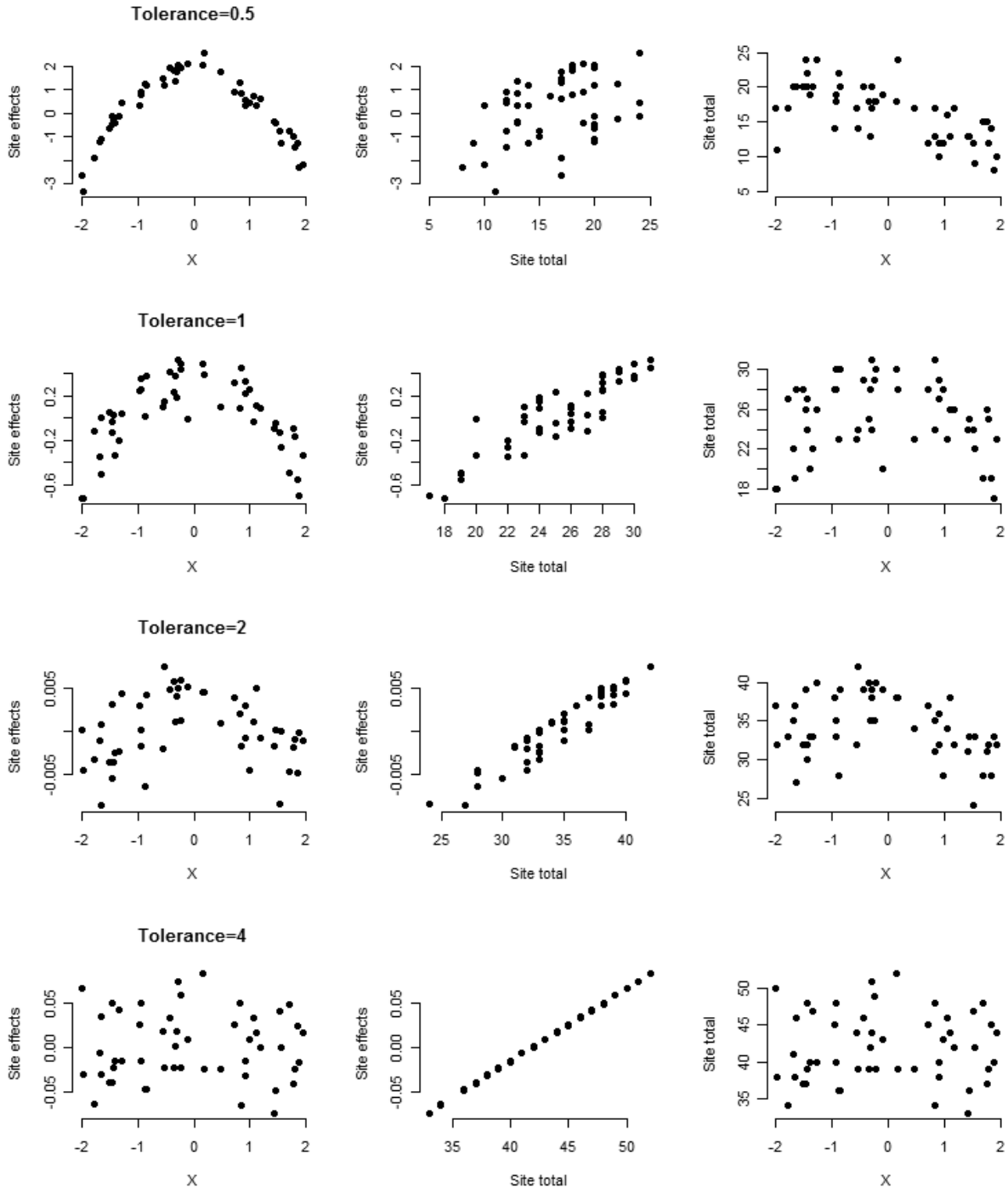

Fig. 2. Diagnostic plot to detect unimodality (4×3). Data simulated from example 1 (Uniform set-up) for four values of tolerances. Tolerance is constant for each of data set. In first column site effects are plotted against the environmental variable for each level of tolerance. In second column sites scores are plotted against the site total (species total per site) and in the third column site total are plotted along the environmental variable.

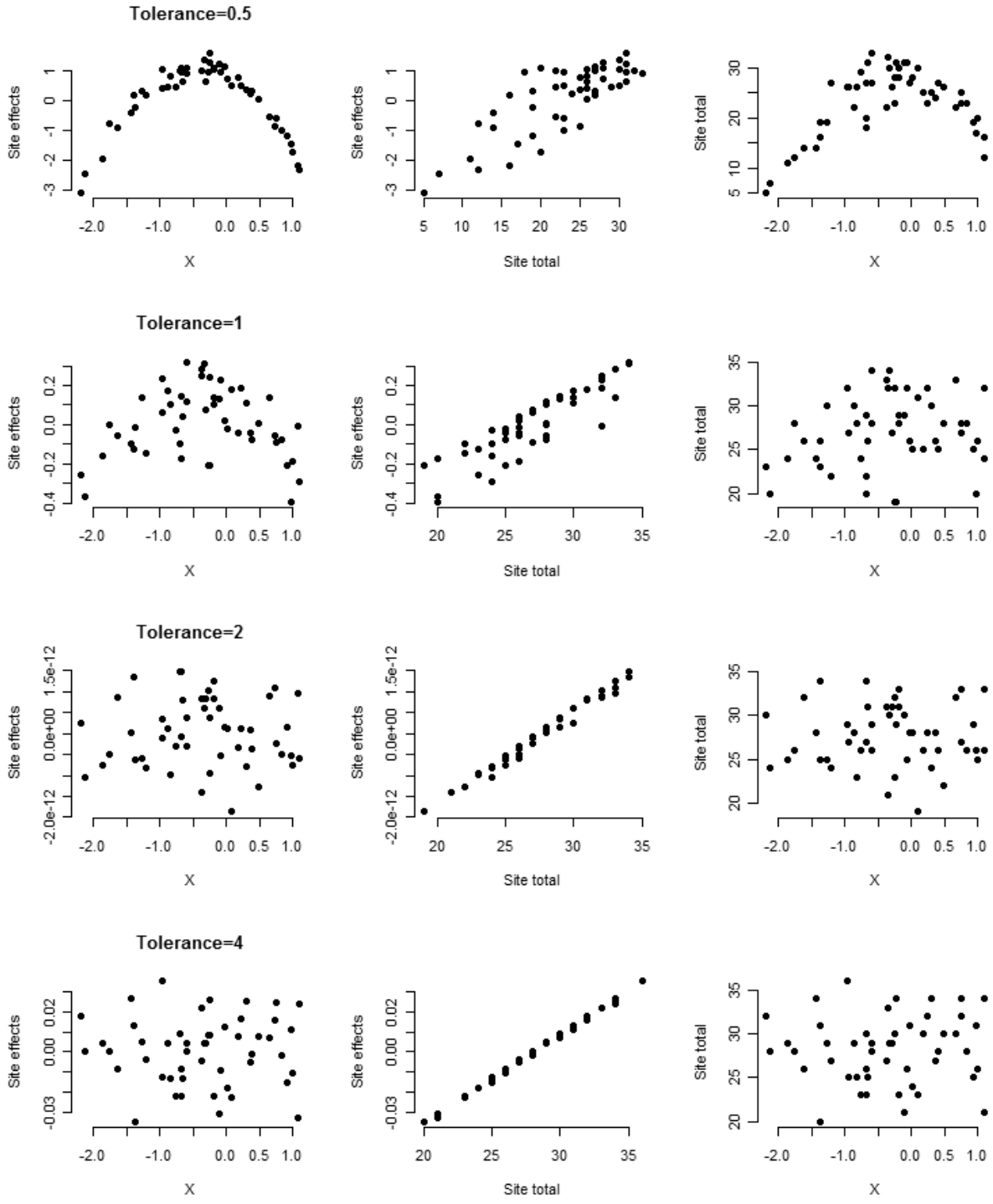

Fig. 3. Diagnostic plot to detect unimodality. Data simulated from **example 2** (Normal set-up) for four values of tolerances. Tolerance is constant for each of data set.

Table 1. Length of gradient and beta diversity of datasets simulated in **example 1** (uniform set-up) and **example 2** (normal set-up) for four levels of tolerance (*t*).

	Uniform setup		Normal setup	
t	Length of gradient	Beta diversity	Length of gradient	Beta diversity
0.5	6.63	4.95	4.81	3.25
1	3.06	2.95	2.77	2.48
2	1.65	1.92	2.04	2.17
4	1.12	1.38	1.73	2.12

Table 2. Coefficient estimate of quadratic term and site total obtained by fitting equation (7) and (8) to the simulated data of **example 1 & example 2**. In parentheses are the corresponding standard errors.

model		$x + x^2$	$x + x^2 + y_{i+}$			
t	σ_γ^2	x^2	x^2		y_{i+}	
Uniform set- up (example 1)						
0.5	2.335	-1.111 (0.037)***	-1.029	(0.012)***	0.105	(0.005)***
1	0.178	-0.232 (0.019) ***	-0.139	(0.002) ***	0.053	(0.001)***
2	0.001	-0.002 (3.7×10^{-3})***	-3.7×10^{-04}	(8.2×10^{-06})***	9.5×10^{-04}	(2.6×10^{-06})***
4	0.010	-0.004 (0.005)	-8.6×10^{-05}	(8.3×10^{-05})***	9.6×10^{-03}	(2.6×10^{-06})***
Normal set-up (example 2)						
0.5	1.623	-1.337 (0.039)***	-1.318	(0.080)***	0.003	(0.011)
1	0.070	-0.136 (0.023)***	-0.089	(0.001)***	0.035	(0.0002)***
2	0.000	0.000 (0.000)	-1.2×10^{-13}	(1.1×10^{-15})***	2.2×10^{-13}	(2.8×10^{-16})***
4	0.005	0.002 (0.003)	-4.1×10^{-4}	(2.6×10^{-6})***	4.3×10^{-03}	(5.7×10^{-7})***

*** p-value < 0.001.

relationship is observed. In example 2 (normal set-up), the plots Figs looks very similar to those of example 1, except that the relationship of the site total with the environmental gradient is more strongly quadratic for $t = 0.5$. In example 1, the site total increases with increasing tolerance but this effect is small in example 2. This might be a side effect of the choice of the distribution of the optima in these examples.

Table 2 shows the relevant coefficients of the regression of the site effect on the environmental variable *x* and its square, with (right) and without (left) the site total in the model. As expected, the coefficient of the squared term is always negative and decreases in size absolute value with increasing tolerance. It may become very close to zero when accounting for the site total. Despite this, the squared term is always significant when the site total is in the model. Without site total, the squared term is not significant for $t = 4$ in both set-ups and for $t = 2$ in the normal set-up. When the data are simulated using a linear model the squared term was not judged significant more often than expected on the basis of Type I error of the test.

With tolerance varying across species with a median of 0.5 and 1, (example 3) the length of

gradient and beta diversity decreases with increasing variation in the tolerance (Table 3). The site effects issued by GLMM still show a quadratic relationship with the environmental variable (Figs. 4 and 5). The quadratic terms decreases in size with increasing coefficient of variation (Table 4), but is significant in all cases.

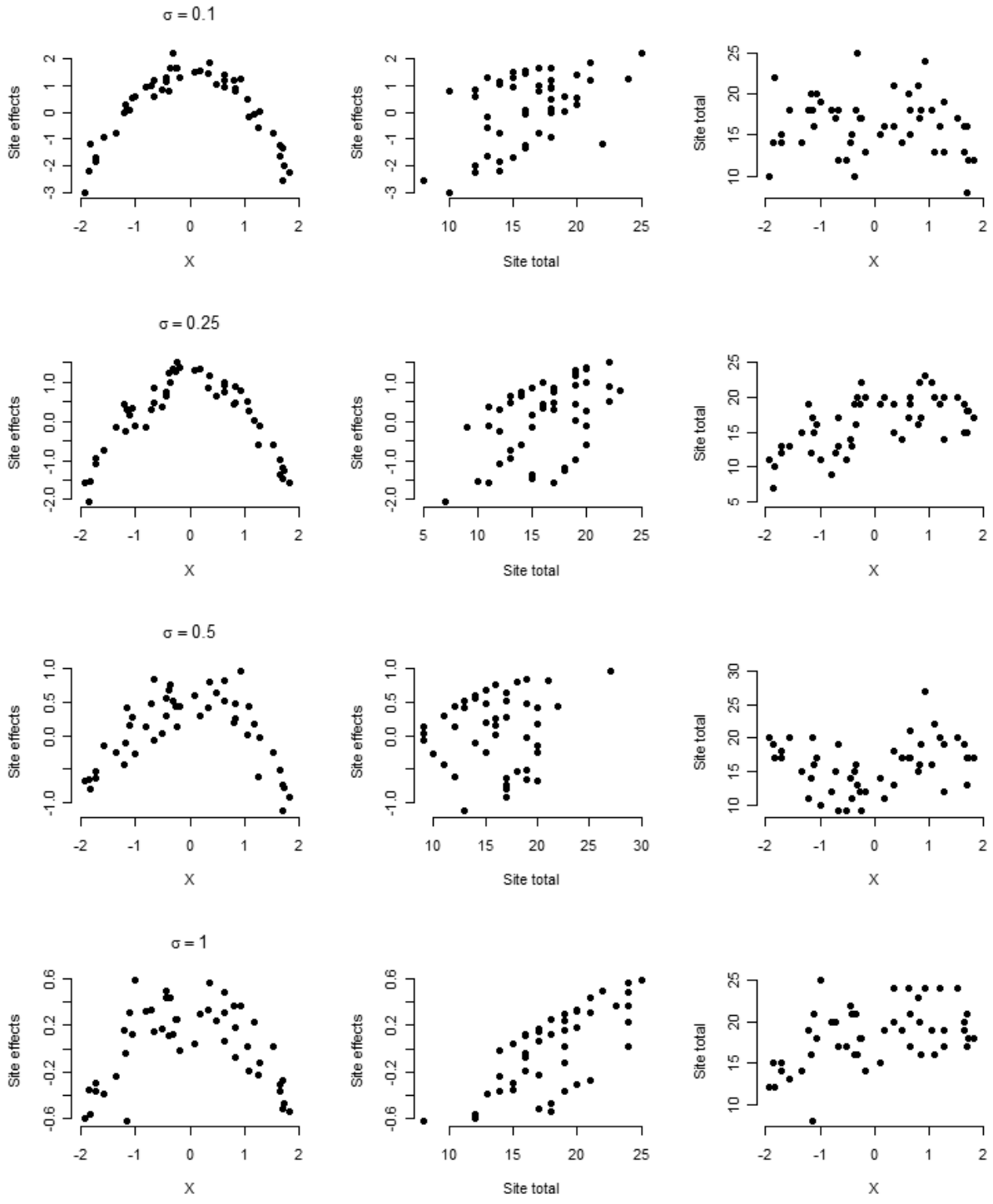

Fig. 4. Varying tolerance $t \sim \text{LogN}(\log(0.5), \sigma)$ for (a) $\sigma = 0.1$, (b) $\sigma = 0.25$ (c)$\sigma = 0.5$ (d)$\sigma = 1$. Data simulated from **example 3A**.

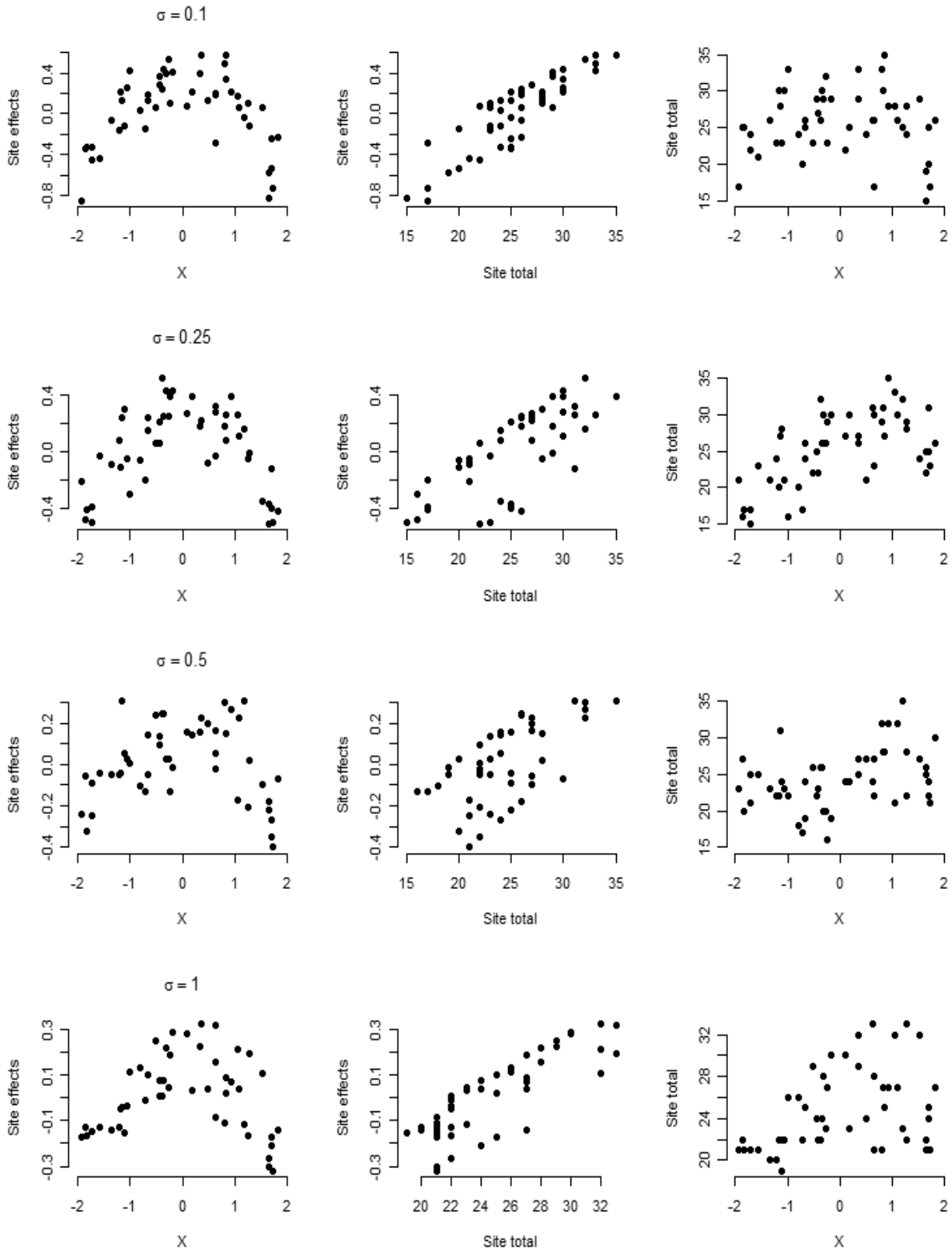

Fig. 5. Varying tolerance $t \sim \text{LogN}(\log(1), \sigma)$ for (a) $\sigma = 0.1$, (b) $\sigma = 0.25$ (c)$\sigma = 0.5$ (d)$\sigma = 1$. Data simulated from **example 3B.**

Table 3. Length of gradient and beta diversity of simulated dataset of **example 3.**

	Lognormal median tolerance=0.5		median tolerance=1	
Sigma	Length of gradient	Beta diversity	Length of gradient	Beta diversity
0.1	6.27	5.03	3.60	2.83
0.25	5.92	5.02	3.04	2.96
0.5	5.42	4.97	3.12	3.07
1	3.60	4.09	2.97	2.81

Table 4. Coefficient estimate of quadratic term and site total obtained by fitting equation (7) and (8) to the simulated data of **example 3**. In parentheses are the corresponding standard errors.

Model		$x + x^2$	$x + x^2 + y_{i+}$	
Sigma	σ_γ^2	x^2	x^2	y_{i+}
Median tolerance =0.5				
0.1	2.110	-1.119 (0.042)***	-1.038 (0.014)***	0.095 (0.005)***
0.25	1.162	-0.805 (0.035)***	-0.702 (0.009)***	0.096 (0.003)***
0.5	0.423	-0.411 (0.034)***	-0.511 (0.008)***	0.078 (0.002)***
1	0.196	-0.241 (0.024)	-0.180 (0.003)***	0.064 (0.001)***
Median tolerance = 1				
0.1	0.201	-0.256 (0.027)***	-0.164 (0.003)***	0.052 (0.001)***
0.25	0.144	-0.209 (0.021)***	-0.128 (0.002)***	0.047 (0.001)***
0.5	0.076	-0.109 (0.017)***	-0.121 (0.002)***	0.036 (0.001)***
1	0.067	-0.109 (0.014)***	-0.068 (0.001)***	0.035 (0.0003)***

*** p-value < 0.001.

Fig. 6 shows the effect of number of species. For smaller ($m = 10$) and larger ($m = 100$) number of species, the site effects remain to show a quadratic relationship with the environmental variable (Fig. 6). The length of gradient and beta diversity decreases as the number of species increases (Table 5). Despite the smaller length of gradient, the quadratic effect is stronger when the number of species is larger (Table 6).

Figs. 7 and 8 shows the linear relationship of between the random slopes (β_j) issued by GLMM and the true optima (u_j). The relation was predicted by equation (3). The relationship is weaker the larger the tolerance. With tolerance varying across species, the relationship continues to hold true surprisingly well (Figs. 9 and 10), except perhaps when the coefficient of variation of the tolerance is large (>100%).

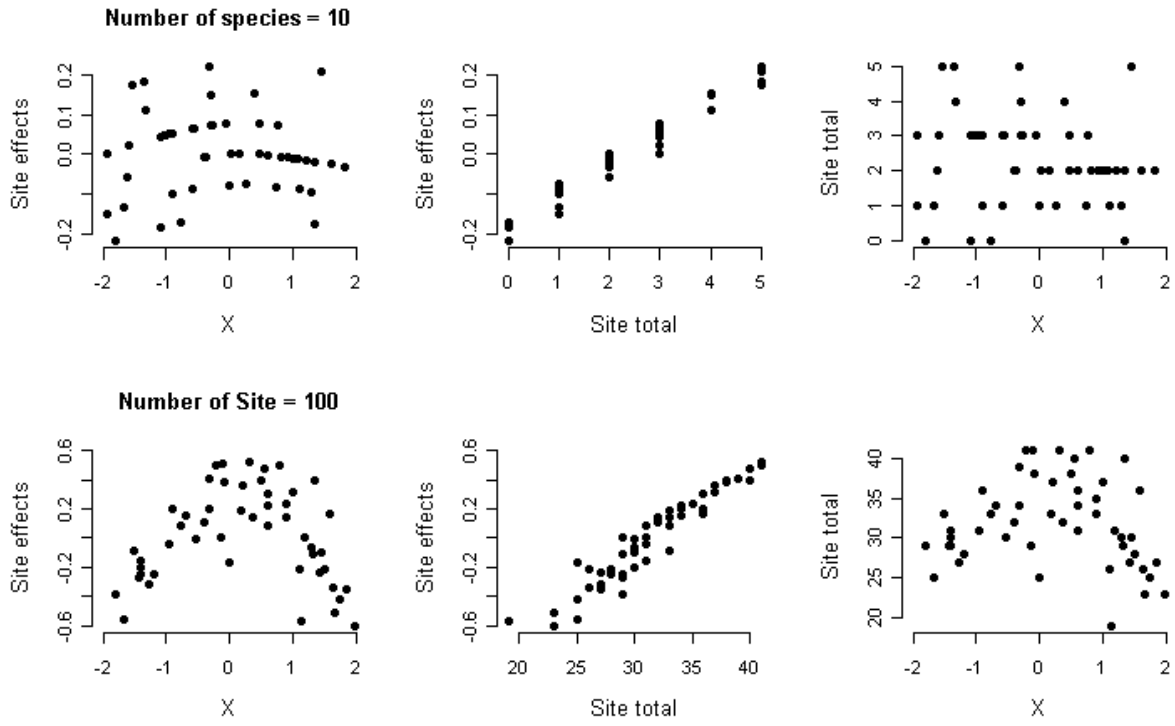

Fig. 6. Data simulated from **example 4** for 10 and 100 species and constant tolerance =1.

Table 5. Length of gradient and beta diversity of simulated dataset of **example 4**

No of Species (*m*)	Length of gradient	Beta diversity
10	4.42	3.31
100	2.75	2.14

Table 6. Coefficient estimate of quadratic term and site total obtained by fitting equation (7) and (8) to the simulated data of **example 4**. In parentheses are the corresponding standard errors and *m* is for number of species.

Model		$x + x^2$	$x + x^2 + y_{i+}$	
m	σ_γ^2	x^2	x^2	y_{i+}
10	0.324	-0.075 (0.017)***	-0.032 (0.001)***	0.225 (0.001)***
100	0.347	-0.363 (0.015)***	-0.190 (0.003)***	0.053 (0.001)***

*** p-value < 0.001.

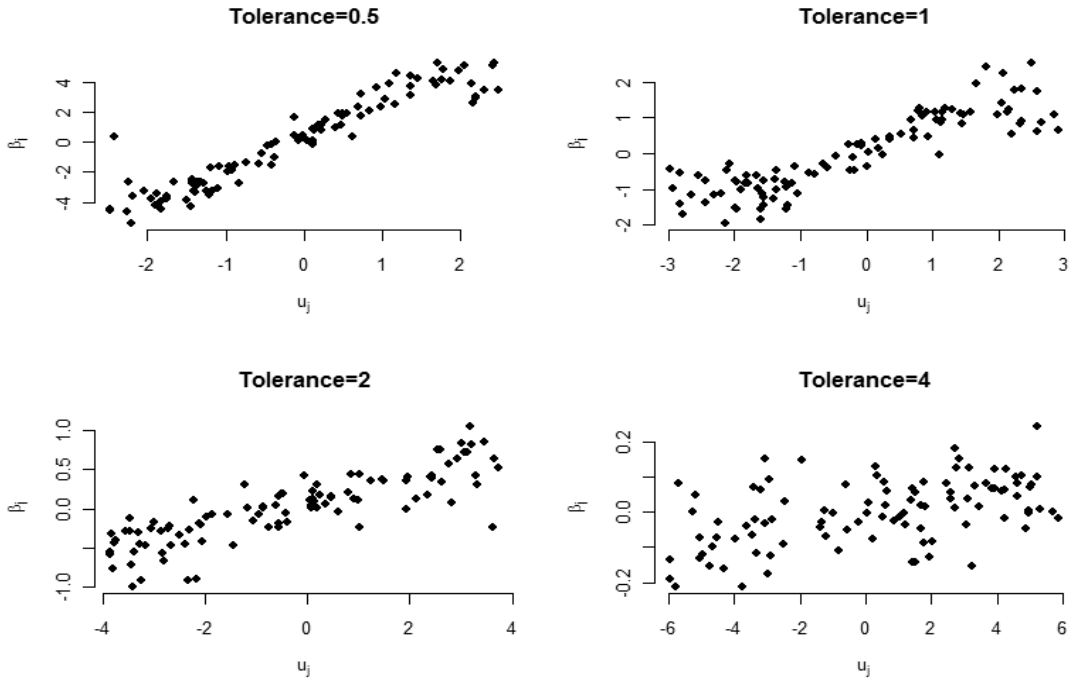

Fig. 7. Plot beta (β_j)vs species optimum (u_j) , data from **example 1.**

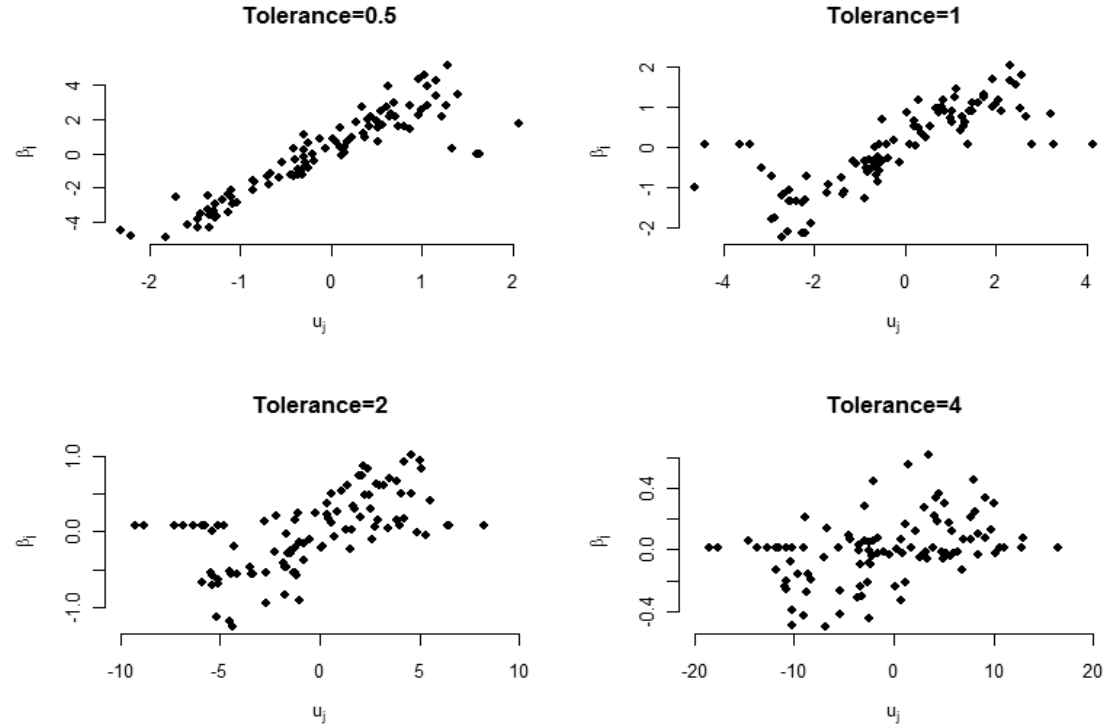

Fig. 8. Plot beta (β_j)vs species optimum (u_j), data from **example 2.**

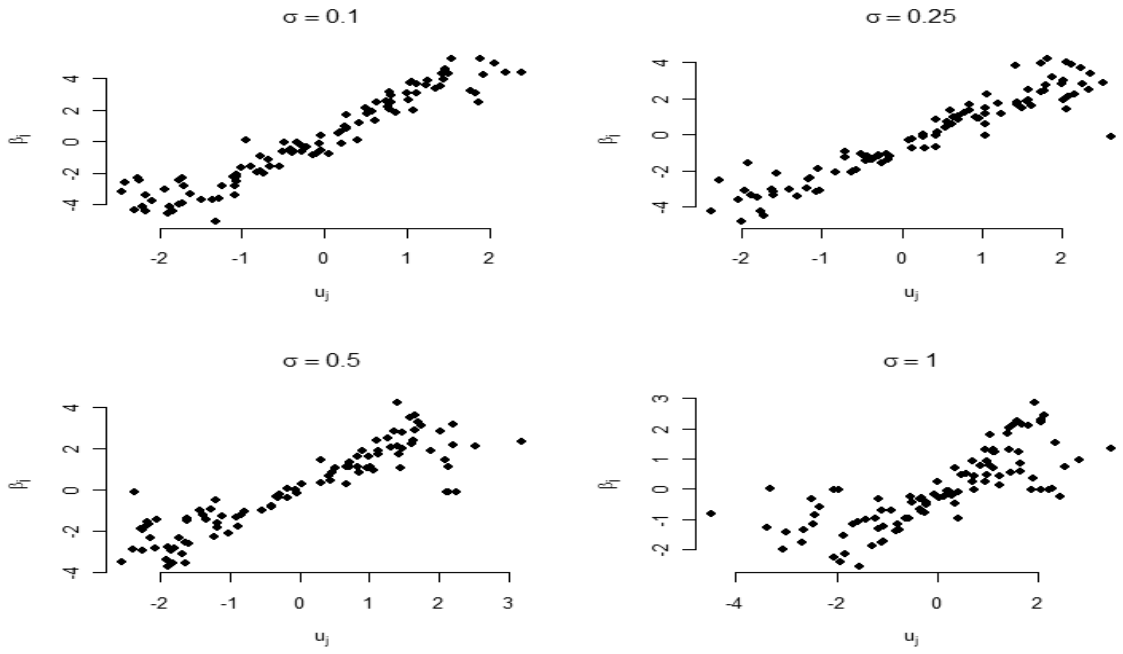

Fig. 9. Plot $\beta_j$vs u_j for **example 3A,** $t\sim$LogN(log(0.5) , σ), Median tolerance=0.5.

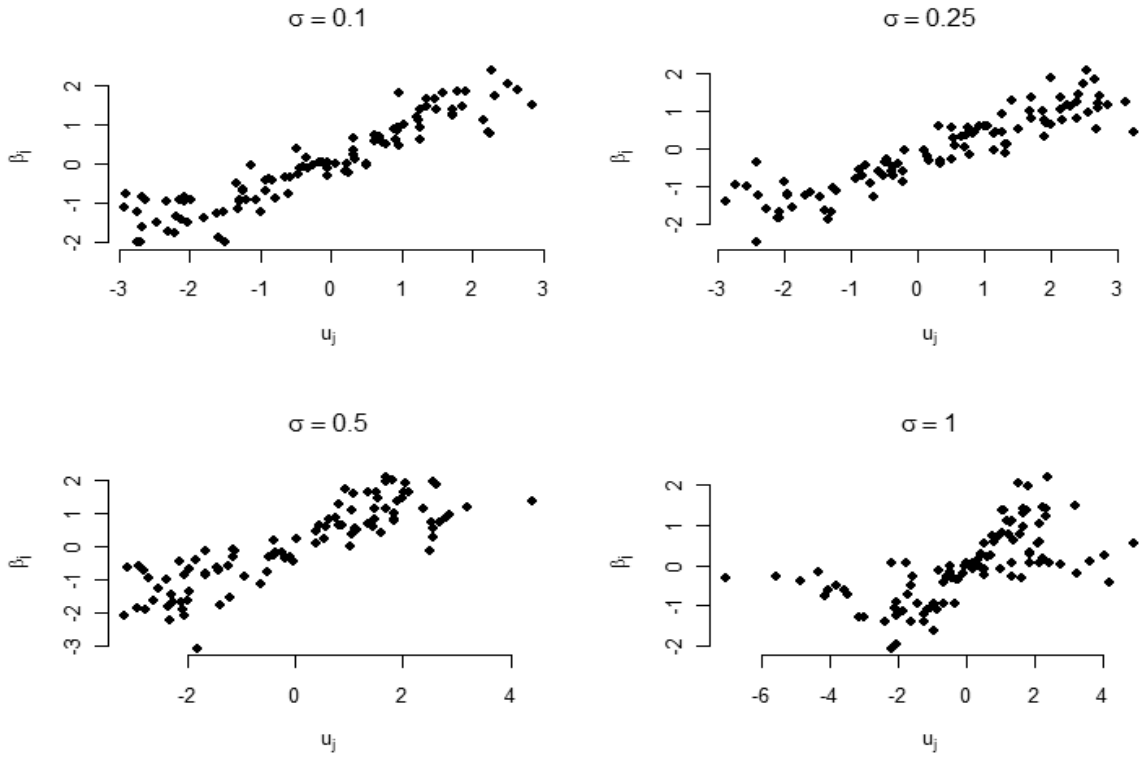

Fig. 10. Plot $\beta_j$vs u_j for **example 3B**. $t\sim$LogN(log(0) , σ), Median tolerance=1.

Real data example

Dune Meadows data

We illustrate the method on the basis of the Dune Meadow data (Jongman et al. 1987). This is a small data set of 28 higher plants in 20 sites in a dune area in the Netherlands. Environmental variables related to soil and management were measured at each site. We fitted a GLMM to the dune meadow data using Moisture as the environmental variable. The estimated site effects show some unimodality when plotted against the environmental variable (Fig. 11). The quadratic term is significant when adjusted for the site total, but without adjustment it is not (Table 7). We conclude that there is some indication for unimodality in this small data set, but the unimodality is not strong.

Phytoplankton data

The data set involve phytoplankton community of 203 lakes located within four climate zones and associated measurements on various environmental variables and morphological species traits of 60 species (Kruk 2010, Jamil et al. in prep)**.** We considered three environmental variables in turn, fitted a GLMM for each and plotted the site effects against the chosen environmental variable. The environmental variables were temperature, chlorophyll-a and a latent variable, that is linear combination of environmental variable $X_L = \mathrm{Chloa} - 0.31 \times \mathrm{Temp} - 0.15 \times \mathrm{Zmix} - 0.25 \times \mathrm{Kd} - 0.02 \times \mathrm{Cond} + 0.05 \times \mathrm{Alk} + 0.01 \times \mathrm{TN} + 0.18 \times \mathrm{TZ}$. The coefficients were estimated using a OpenBUGS program that specified a Bayesian variable selection method.

Both the graphical test and the statistical test confirm unimodality (Fig. 12, Table 7). Fig. 13 shows the relationship of between the random slopes (β_j) with respect to chlorophyll-a issued by GLMM and the optima (u_j) on the chlorophyll-a gradient as obtained from a fit of the unimodal model of Eq. 1using OpenBUGS. The species with low values for the optimum received similar values for the slope, analogously to Figs. 7-10, and thus cannot be properly ranked on the basis of the slopes only, but otherwise there is a good agreement.

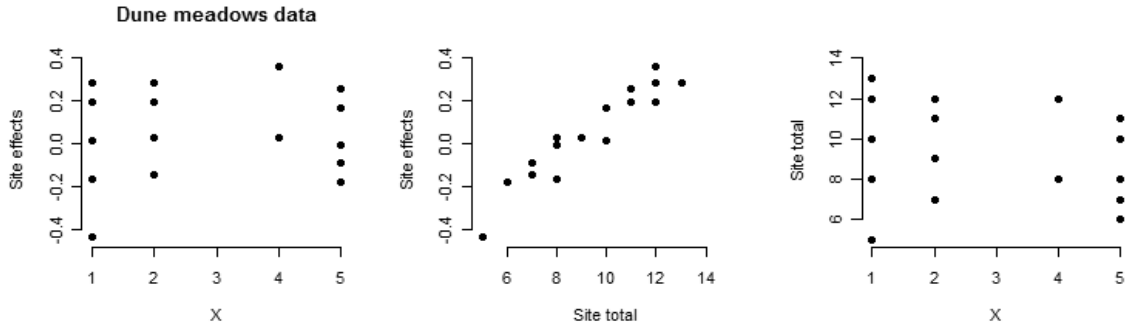

Fig. 11. Diagnostic plot from Dune Meadows data where Moisture is an environmental gradient.

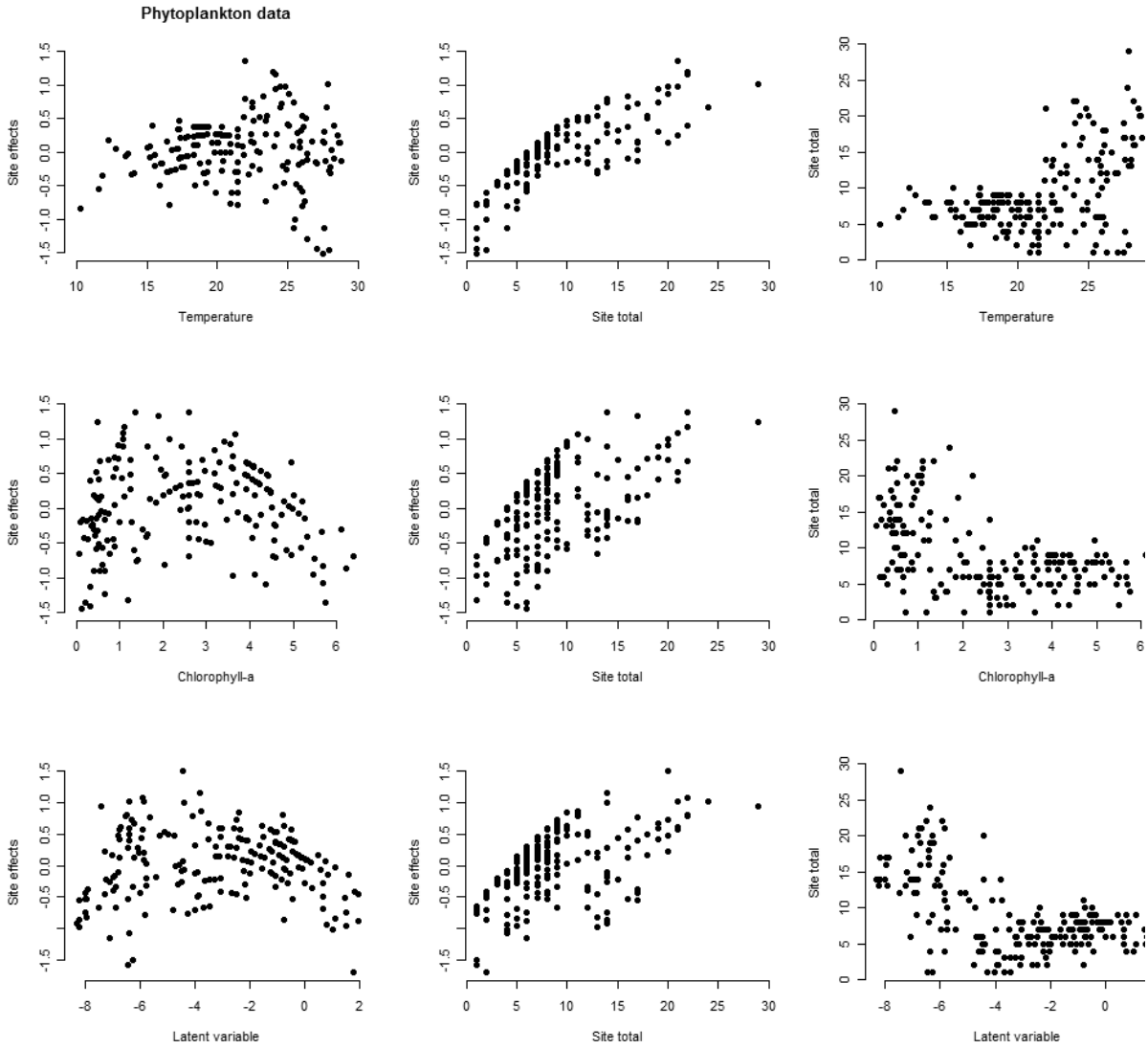

Fig. 12 Diagnostic plot for Phytoplankton data. Three rows are for temperature, chlorophyll-a, and latent variable.

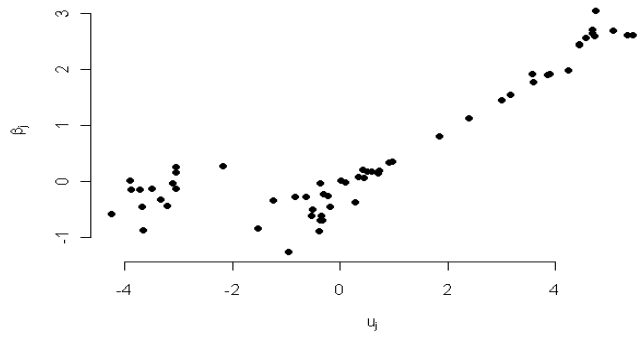

Fig. 13. Plot of the slope estimates (β_j) of the GLMM against the fitted optimum (u_j) obtained with unimodal Gaussian model.

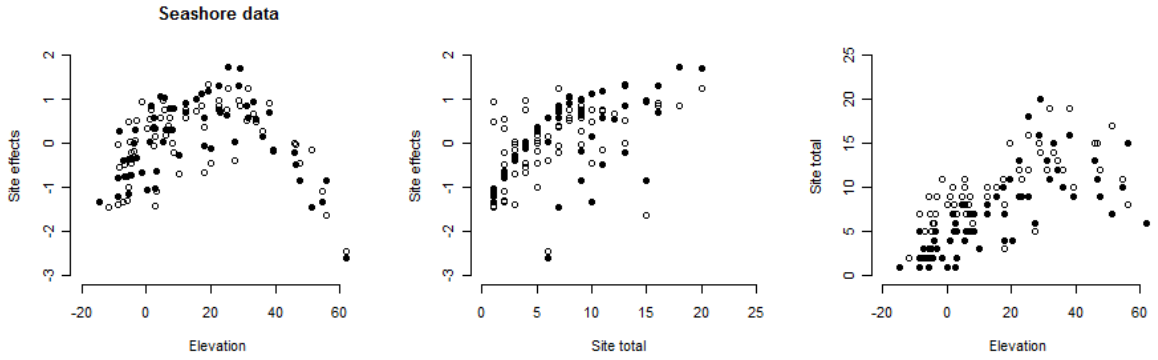

Fig. 14. Diagnostic plot for Sea-shore data for the two years (circle (●) for the year 1978 and circle (○) for the year 1984) where elevation is an environmental gradient

Sea-shore data

This data set is about the vegetation of the rising seashore on the island Skabbholmen in the Stockholm archipelago, eastern central Sweden (Cramer and Hytteborn 1987) and is part of the Canoco package (ter Braak and Smilauer 1998). The data set consists of 63 sites sampled in both 1978 and 1984 and contains 68 species. We fitted a GLMM to the sea-shore data for the years 1978 and 1984 separately and plotted the site effects against the environmental variable Elevation for both years. The estimated site effects show unimodality when plotted against the environmental variable (Fig. 14, Table 7). Both the graphical test and the statistical test confirm unimodality.

Table 7. Coefficient estimate of quadratic term and site total obtained by fitting equation (7) and (8) to the dune meadow data and phytoplankton data. In parentheses are the corresponding standard errors. In parentheses are the corresponding standard errors

Model		$x + x^2$	$x + x^2 + y_{i+}$	
	σ_γ^2	x^2	x^2	y_{i+}
Dune meadows data				
Moisture	0.135	-0.053 (0.034)	-0.020 (0.001) ***	0.088 (0.001)***
Phytoplankton data				
Temprature	0.375	-0.005 (0.002) ***	-0.012 (0.0003)***	0.103 (0.001)***
chlorophyll-a	0.606	-0.114 (0.016) ***	-0.182 (0.005) ***	0.124 (0.002)***
Latent variable	0.463	-0.035 (0.005) ***	-0.064 (0.001) ***	0.117 (0.002)***
Sea-shore data				
Elevation-1978	1.113	-0.002 (0.0002)***	-0.001 (0.0001)***	0.158 (0.010)***
-1984	0.973	-0.002(0.0002)***	-0.001 (7.0×10^{-5})***	0.171 (0.008)***

*** p-value < 0.001.

Conclusion

To our knowledge, explicit testing of unimodality in species response along an environmental gradient, without fitting a unimodal model, has not been done before. Walker and Jackson (2011) used a latent variable approach to test for unimodality. We tried their approach to the phytoplankton data, but failed to get an answer because the program for fitting the quadratic model crashed. In this paper, we take a simpler approach and studied the suitability of GLMM for detecting the unimodality of species response along an environmental gradient and suggested a graphical tool and a statistical test for testing unimodality. There is an indication for unimodality when site effects show quadratic relationship with the environmental gradient. The test can make even stronger by adjusting the relationship with the site total (the number of species in a site).

As an alternative to our approach, we could explicitly add the square of the environmental variable as a fixed effect term to the GLMM of equation (4), yielding

$$\text{logit}(p_{ij}) = \alpha_j + \beta_{1j} x_i + \beta_2 x_i^2 + \gamma_i \qquad (14)$$

and judge the significance of the addition by a one-sided test on β_2 ($H_0: \beta_2 = 0$ versus $H_1: \beta_2 < 0$). This approach is presumably even more powerful, but necessitates the fit of an extra model. The model assumes constant tolerance for all species curves (as does equation (4)) and can be rewritten as

$$\text{logit}(p_{ij}) = a_j - \frac{(x_i - u_j)^2}{2t^2} + \tilde{\gamma}_i \qquad (15)$$

with $t = 1/\sqrt{-2\beta_2}$, $u_j = t^2 \beta_j$, $a_j = \alpha_j + \frac{1}{2t^2} u_j^2$, and $\tilde{\gamma}_i = \gamma_i + \frac{1}{2t^2} x_i^2$. (16)

To test the assumption of equi-tolerance, we can go one step further and add the squared term x^2 also as a random (species-dependent) component to equation (14) and test the significance of this extra variance component. In the R package lme4 (Bates et al. 2011), the two models to compare are (with xx = x^2)

```
lmer (y ~ 1+ x + xx + (1+ x | sp) + (1| site)
          family=binomial(link="logit") ,data)
```

and

```
lmer (y ~ 1+ x + xx + (1+ x + xx| sp) + (1| site)
          family=binomial(link="logit") ,data).
```

A GLMM is, of course, a linear model. This paper shows that, despite this fact, it can be used to detect unimodality and to fit unimodal data, with the provision that the differences in niche widths among species is not too large (Fig. 10) and even this assumption can be tested within the GLMM framework. The application scope of GLMM in ecology is thus much broader than one might think at first glance.

Chapter 5

Selection properties of Type-II maximum likelihood (empirical Bayes) in linear models with individual variance components for predictors

Tahira Jamil, Cajo J.F. ter Braak
(Submitted to Pattern Recognition Letters)

Abstract

Maximum Likelihood (ML) in the linear model overfits when the number of predictors (*M*) exceeds the number of objects (*N*). One of the possible solution is the Relevance vector machine (RVM) which is a form of automatic relevance detection and has gained popularity in the pattern recognition machine learning community by the famous textbook of Bishop (2006). RVM assigns individual precisions to weights of predictors which are then estimated by maximizing the marginal likelihood (Type-II ML or empirical Bayes). We investigated the selection properties of RVM both analytically and by experiments in a regression setting.

We show analytically that RVM selects predictors when the absolute z-ratio (|least squares estimate|/standard error) exceeds 1 in the case of orthogonal predictors and, for $M = 2$, that this still holds true for correlated predictors when the other z-ratio is large. RVM selects the stronger of two highly correlated predictors. In experiments with real and simulated data, RVM is outcompeted by other popular regularization methods (LASSO and/or PLS) in terms of the prediction performance. We conclude that Type-II ML is not the general answer in high dimensional prediction problems.

In extensions of RVM to obtain stronger selection, improper priors (based on the inverse gamma family) have been assigned to the inverse precisions (variances) with parameters estimated by penalized marginal likelihood. We critically assess this approach and suggest a proper variance prior related to the Beta distribution which gives similar selection and shrinkage properties and allows a fully Bayesian treatment.

Keywords: Automatic relevance detection; Empirical Bayes; Lasso; Sparse model; Type-II maximum likelihood; Relevance vector machine

Introduction

Maximum likelihood (ML) or least squares (LS) can lead to sever over-fitting and poor estimation, when the number of predictors or basis functions (M) is large as compared to data size (N) *i.e.*, $M \geq$ N. Regularization or shrinkage estimation can improve an estimate and regularize an ill-posed problem (Bishop 2006). This involves adding a penalty term to the error function in order to discourage parameters from reaching large values. In a linear model the modified error function takes the form

$$\text{RSS} + \lambda \sum_{m=1}^{M} |w_j|^q \text{ for } q \geq 0, \quad (1)$$

where RSS is the residual sum of squares, $\mathbf{w} = (w_1,.., w_M)^T$ is the parameter vector containing the weights (regression coefficients) for the predictors, and $\lambda \geq 0$ is a complexity parameter that controls the amount of regularization. For q=2 we have ridge regression (RR) (Hoerl and Kennard 1970) which proportionally shrinks estimates of $\{w_j\}$ to zero, but does not produce a sparse solution. In neural networks this is known as weight decay. For q=1 we have the LASSO (least absolute shrinkage and selection operator) (Tibshirani 1996) which also shrinks the coefficients towards zero but also puts some coefficients exactly to zero, and therefore performs variable selection (Tibshirani 1996, Efron et al. 2004). The optimal choice for λ in penalized likelihood is often based on cross validation.

Most regularization methods have a Bayesian interpretation as giving the maximum a posterior (MAP) mode for a given prior distribution for the parameters. The prior in RR is Gaussian and in LASSO it is double exponential. The equivalence of MAP with the shrinkage estimate does not mean that the Bayesian framework is simply a re-interpretation of classical methods. The distinguishing element of Bayesian inference is marginalization. By marginalizing over **w** we obtain a marginal likelihood, also known as the Type-II likelihood or the evidence function (Bishop 2006). The parameter λ can then be obtained by maximizing this function, *i.e.* by Type-II maximum likelihood, and then **w** is obtained for this value of λ. This procedure is also known as empirical Bayes and automatic relevance determination (MacKay 1992, Neal 1996). A fully Bayesian approach would also require a prior for the hyperparameter λ and marginalization over λ.

Tipping (2001) created the relevance vector machine (RVM) as a sparse kernel technique build upon a linear model with $M = N$. In RVM each weight w_j is assigned an independent Gaussian prior with an individual precision, resulting in M hyperparameters which are all precisions (or their inverse, variances). Tipping (2001) considered assigning a Gamma prior to the precisions, but

eventually focussed on a uniform prior for which maximization of the posterior reduces to maximization of the marginal likelihood, also called the type II likelihood (Tipping 2001, Bishop 2006). By maximizing the Type-II likelihood with respect to all M hyperparameters many precisions go to infinity (Faul and Tipping 2002), so creating a sparse model as each infinite precision effectively eliminates the corresponding predictor from the model. Tipping and Faul (2003) developed a fast sequential algorithm for this. RVM has found wide-spread application with 705 citations in the Web of Science as of July 2011, also outside the kernel world (Li et al. 2002, Rogers and Girolami 2005) and found general exposure through the exposition in Bishop (2006). However, little is known about the properties of RVM. With (hyper)parameters on the edge of the permissible region, general asymptotic theory for maximum (marginal) likelihood does not apply.

This paper studies the selection and shrinkage properties of RVM in the un-kernelised regression setting (Bishop 2006: section 7.2). We found it easier to work with variance rather than precision, because a predictor drops from the model when its variance component is zero, which is easier to work with than with infinite precision. As Bishop (2006), we phrase and study RVM outside its kernel context as a Type-II maximum likelihood approach to the linear model with individual variance components for the predictors. We first state the model and rewrite the marginal likelihood in a form that uses inner product matrices of size $M \times M$ rather than $N \times N$. We then obtain an analytical expression for the selection and shrinkage properties of RVM in the special case that the predictors are orthonormal and the error variance is known. The main result here is that RVM drops a predictor from the model if and only if its z-ratio (least-squares estimate of the weight divided by its standard error of estimate) is less than 1 in absolute value. RVM is thus very tolerant in allowing predictors to stay in the model. In practice, particularly in a kernel context and always when $M > \mathrm{N}$, predictors are not orthogonal and regularization methods tend to behave very different in the presence of correlation. For example, if the two predictors are highly correlated, LASSO selects one, whereas ridge, elastic net (Zou and Hastie 2005) and PLS (Frank and Friedman 1993) select both. Tibshirani (1996) gave analytical expressions for the two correlated predictors case for the LASSO. In section 4 we attempt similarly for RVM and arrive at analytical expressions for when RVM selects neither, one or both predictors. The main conclusion from these expressions is again that RVM is very tolerant in allowing predictors to stay in the model. In section 5 we compare RVM on simulated and real data for a range of M/N ratios with LASSO and Partial Least Squares (PLS), which is a shrinkage method based on latent variables that is very popular in chemometrics (Wold et al. 2001). We conclude with a discussion of the RVM and its extensions in relation to fully Bayesian approaches.

RVM as sparse Bayesian linear regression

RVM for regression is a linear model with a prior that results in a sparse solution (Bishop 2006). The model for real-valued target variable *t*, given an input vector **x**, takes the form

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon \tag{2}$$

where **w** is a vector of *M* parameters and ϵ is a white noise term that is Gaussian distributed with zero mean and variance σ^2, which we will assume known.The regression function $\boldsymbol{y}(\mathbf{x}, \mathbf{w})$ is then defined as the linear model

$$y(\mathbf{x}, \mathbf{w}) = \sum_{m=1}^{M} w_m \phi_m(\mathbf{x}) \tag{3}$$

with fixed nonlinear basis functions $\phi_m(\mathbf{x})$. For ease of presentation we ignore the constant term representing bias as it can be dealt with by centring the target variable. For given set of *N* independent observations of the target *t* and input vector **x**, the data likelihood function of the target vector $\mathbf{t} = (t_1, \dots, t_N)^T$ for given input vectors $\{\mathbf{x}_i\}_{i=1,\dots,N}$ is

$$p(\mathbf{t}|\mathbf{w}, \sigma^2) = (2\pi)^{-N/2} \sigma^{-N} \prod_{i=1}^{N} \exp\left(-\frac{1}{2\sigma^2}\left(t_i - y(\mathbf{x}_i, \mathbf{w})\right)^2\right). \tag{4}$$

To make it a Bayesian model we need to specify a prior for the parameter **w**. In RVM, each parameter w_m is an independent zero mean Gaussian with a separate variance parameter α_m, giving

$$p(\boldsymbol{w}|\boldsymbol{\alpha}) = (2\pi)^{-M/2} \prod_{m=1}^{M} \alpha_m^{-1/2} \exp\left(-\frac{w_m^2}{2\alpha_m}\right) \tag{5}$$

where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_M)^T$ is the vector of hyperparameters, which are in our notation *not* precisions but variances. These *M* independent hyperparameters control the strength of the prior over its associated weight and this form of prior is responsible for the sparsity properties of the model (Tipping 2001).

In Type-II maximum likelihood (Berger 1985), also known as empirical Bayes or evidence approximation (MacKay 1992), an estimate $\widehat{\boldsymbol{\alpha}}$ is obtained by maximizing the marginal likelihood $p(\boldsymbol{t}|\boldsymbol{\alpha}, \sigma^2)$ over $\boldsymbol{\alpha}$, which is then plugged into posteriori density $p(\mathbf{w}|\mathbf{t}, \hat{\alpha}, \sigma^2)$, which is a multivariate normal, the mean which is taken as the shrinkage estimate $\widetilde{\boldsymbol{w}}$. The marginal likelihood requires integration over $\boldsymbol{w}$, giving the multivariate normal density (Bishop 2006)

$$L(\boldsymbol{\alpha}) = p(\mathbf{t}|\boldsymbol{\alpha}, \sigma^2) = \int_{-\infty}^{\infty} p(\mathbf{t}|\mathbf{w}, \sigma^2) p(\boldsymbol{w}|\boldsymbol{\alpha}) d\mathbf{w} = (2\pi)^{-N/2} |\mathbf{C}|^{-1/2} \exp\left(-\frac{1}{2}\mathbf{t}^T \mathbf{C}^{-1} \mathbf{t}\right), \tag{6}$$

where $\mathbf{C} = \sigma^2 \mathrm{I} + \mathbf{\Phi A \Phi}^T$ with $\mathbf{\Phi}$ the $N \times M$ design matrix, of which the ith row is $(\varphi_1(\mathbf{x}_i), \ldots, \varphi_m(\mathbf{x}_i))^T$ and $\mathbf{A} = \mathrm{diag}(\alpha_1, \ldots, \alpha_M)$.

Our goal is now to maximize (6) with respect to the hyperparameters $\boldsymbol{\alpha}$. At this point we deviate from Bishop (2006) and convert the inverse and determinant of the $N \times N$ matrix $\mathbf{C}$ using the matrix inversion and determinant lemma or Woodbury formula (Golub and van Loan 1989) into forms using $M \times M$ matrices. On deleting terms that do not depend on $\boldsymbol{\alpha}$, we obtain (Appendix A)

$$L(\boldsymbol{\alpha}) \propto |\mathbf{I} + \sigma^{-2}\mathbf{\Phi}^T\mathbf{\Phi A}|^{-1/2} \exp\left(\frac{1}{2\sigma^2}\mathbf{t}^T\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi} + \sigma^2\mathbf{A}^{-1})^{-1}\mathbf{\Phi}^T\mathbf{t}\right). \quad (7)$$

This marginal likelihood has a form equivalent to the posterior distribution of the variance component in a hierarchical linear model or random model (O'Hagan and Forster 2004, ter Braak 2006). The study of the selection properties of RVM is equivalent to the study of the conditions under which hyperparameters (α-values) become zero. We do this by setting the derivative of (7) with respect to $\boldsymbol{\alpha}$ to zero, solving the resulting equation for $\boldsymbol{\alpha}$, checking that this is a maximum and checking whether the obtained $\hat{\boldsymbol{\alpha}}$ has some zero elements.

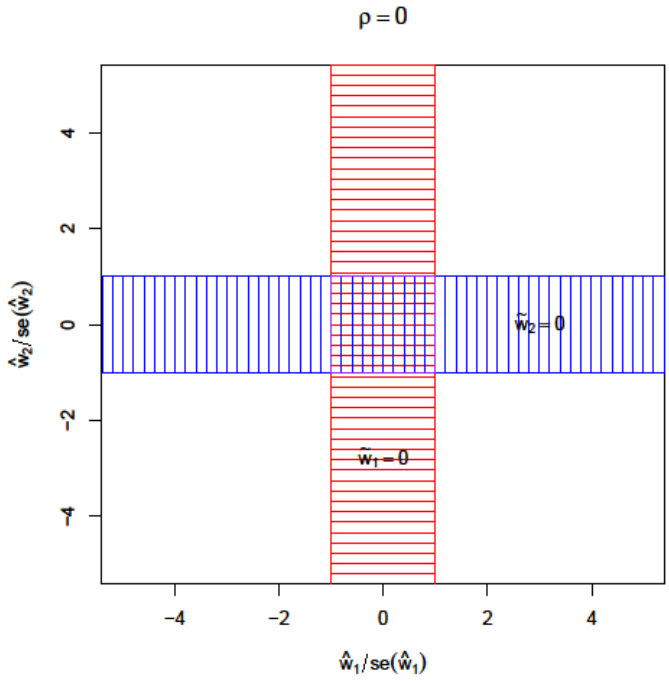

Fig. 1. Two uncorrelated predictor case: regions in terms of the z-ratio (estimate/standard error) where the RVM estimate of the weights and associated variance components are exactly zero. In these regions the corresponding predictor(s) can be pruned from the model.

Orthonormal predictors

In this section we study the selection properties of RVM in the special case that the predictors are orthogonal, *i.e.* $\mathbf{\Phi^T\Phi}$ is a diagonal matrix. In this case the marginal likelihood (7) decomposes as a product of individual likelihoods $L(\alpha_m)$ with (Appendix B)

$$L(\alpha_m) \propto (1 + v_m^{-1}\alpha_m)^{-1/2}\exp\left(\frac{\hat{w}_m^2 v_m^{-2}\alpha_m}{2(1+v_m^{-1}\alpha_m)}\right) \quad (8)$$

where $\hat{w}_m = \boldsymbol{\phi}_m^T\mathbf{t}/\boldsymbol{\phi}_m^T\boldsymbol{\phi}_m$, the least-squares estimate, $v_m = \sigma^2/\boldsymbol{\phi}_m^T\boldsymbol{\phi}_m$, the variance of $\hat{w}_m$, and $\boldsymbol{\phi}_m$ is the *m*th column of $\mathbf{\Phi}$. The variance component α_m that maximizes (8) is

$$\hat{\alpha}_m = (\hat{w}_m^2 - v_m)_+, \quad (9)$$

where $(.)_+$ is the positive part operator, defined as $(a)_+ = a$ if $a > 0$ and 0 otherwise. In the orthogonal predictor case, RVM thus leads to soft thresholding (Donoho and Johnstone 1994, Donoho 1995) of the variance component, whereas LASSO does this for the weights (Tibshirani 1996). Also observe that $\hat{\alpha}_m = 0$, iff $\hat{w}_m^2 \leq v_m$ or, equivalently, |z-ratio| $\equiv |\hat{w}_m/se(\hat{w}_m)| \leq 1$ where *se*(.) is the standard error of estimate. The elements of the shrinkage estimate $\tilde{\boldsymbol{w}}$ for which the z-ratio is smaller than 1, are thus zero. The corresponding predictors can thus be pruned. Fig. 1 displays the result for the case of two uncorrelated predictors.

Two correlated predictors

We now consider the case with two correlated predictors and assume they are rescaled such that $\boldsymbol{\phi}_1^T\boldsymbol{\phi}_1 = \boldsymbol{\phi}_2^T\boldsymbol{\phi}_2 = 1$ and $\boldsymbol{\phi}_1^T\boldsymbol{\phi}_2 = \boldsymbol{\phi}_2^T\boldsymbol{\phi}_1 = \rho$. In this case, $\mathbf{A}$ is a 2×2 diagonal matrix with diagonal elements α_1 and α_2 which are linearly changed by the rescaling. The dependence of the marginal likelihood $L(\boldsymbol{\alpha})$ on σ^2 can be removed by transformation to variance ratios $\boldsymbol{\gamma} = \boldsymbol{\alpha}/\sigma^2$ and by defining $c = \boldsymbol{\phi}_1^T\mathbf{t}/\sigma$, $d = \boldsymbol{\phi}_2^T\mathbf{t}/\sigma$. The maximum is invariant under these transformations. Note that c is the simple z-ratio, that is the z-ratio in least-squares simple regression with single predictor $\boldsymbol{\phi}_1$, and the same holds for d and $\boldsymbol{\phi}_2$. On using Mathematica, differentiating $L(\alpha_1, \alpha_2)$ with respect to γ_1 and γ_2 and setting the derivatives equal to zero gives (Appendix C)

$$\begin{bmatrix}\hat{\gamma}_1\\ \hat{\gamma}_2\end{bmatrix} = \begin{bmatrix}\frac{\left(c+\hat{\gamma}_2(c-\rho d)\right)^2-(1+\hat{\gamma}_2)\left(1+\hat{\gamma}_2(1-\rho^2)\right)}{(1+\hat{\gamma}_2(1-\rho^2))^2}\\ \frac{\left(d+\hat{\gamma}_1(d-\rho c)\right)^2-(1+\hat{\gamma}_1)\left(1+\hat{\gamma}_1(1-\rho^2)\right)}{\left(1+\hat{\gamma}_1(1-\rho^2)\right)^2}\end{bmatrix}. \quad (10)$$

This represents a maximum if $\hat{\gamma}_1 \geq 0$ and $\hat{\gamma}_2 \geq 0$. From (10), if $\hat{\gamma}_1 = 0$, then $\hat{\gamma}_2 = (d^2 - 1)_+$ and it should hold that

$$\left(c + \hat{\gamma}_2(c - \rho d)\right)^2 \leq (1 + \hat{\gamma}_2)\left(1 + \hat{\gamma}_2(1 - \rho^2)\right). \quad (11)$$

Inserting $\hat{\gamma}_2 = (d^2 - 1)_+$ in (11) and solving for c gives upper and lower bounds

$$c = \rho d(1 - d^{-2}) \pm \sqrt{1 - \rho^2(1 - d^{-2})} \quad (12)$$

Subject to $\hat{\gamma}_2 > 0$, values of c within the bounds of (12) give $\hat{\gamma}_1 = 0$, and consequently $\tilde{w}_1 = 0$. Interchanging the roles of $\hat{\gamma}_1$ and $\hat{\gamma}_2$ and thus c and d yields similar for bounds of d such that $\hat{\gamma}_2 = 0$, and consequently $\tilde{w}_2 = 0$, now subject to $\hat{\gamma}_1 > 0$. If $\hat{\gamma}_1 = 0$, then $\hat{\gamma}_2 = (d^2 - 1)_+$, and if $\hat{\gamma}_2 = 0$, then $\hat{\gamma}_1 = (c^2 - 1)_+$, so that both variance ratios are zero if both $|c|$ and $|d|$ are smaller than 1. The bounds are a function of c, d and ρ^2.

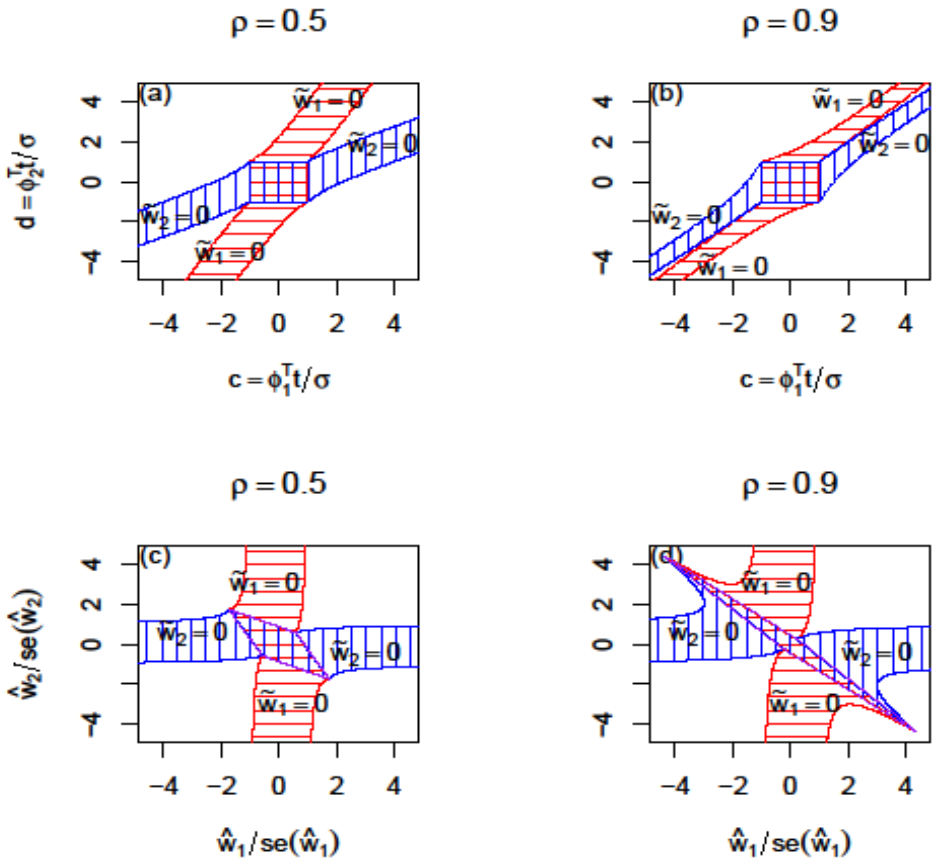

Fig. 2. Two correlated predictor case: regions in terms of the simple) z-ratio (a,b) and multiple z-ratio (c,d) where the RVM estimate of the weights and associated variance ratios are exactly zero. In these regions RVM prunes the corresponding predictor(s) from the model. The simple z-ratio is based on least-squares with a single predictor, the (multiple) z-ratio on least-squares with two predictors.

Figs. 2a,b shows these bounds for ρ = 0.5 and 0.9 in the (*c*,*d*)-plane and the resulting regions where none, one or both weights are exactly zero. Figures for the corresponding negative values of ρ differ only in rotation over 90° and shading.

Whereas for ρ = 0, $\widetilde{w}_1 = 0$ if the simple z-ratio *c* is less than 1 in absolute value ($|c| < 1$), no such simple rule exists for ρ ≠ 0. The interval of *c*-values for which the first weight is exactly zero depends on *d*, as shown in Figs. 2a,b. For example, for ρ = 0.9 then still $\widetilde{w}_1 = 0$ for $|c| < 1$ if d = 1, but if d = 4, then $\widetilde{w}_1 = 0$ if $2.88 < c < 3.87$ (Fig. 2b). The simple z-ratio alone thus says little about the nullity of the first weight estimate. We need both *c* and *d*.

Figs. 2c,d shows the same bounds in terms of the (multiple) z-ratio's, $\widehat{w}_m/se(\widehat{w}_m)$, m = 1,2, *i.e.* $\widehat{\boldsymbol{w}} = \left(\boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\boldsymbol{T}}\mathbf{y}$ and $se(\widehat{w}_m)^2$ is a diagonal element of $\sigma^2\left(\boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{\Phi}\right)^{-1}$, so that, in the two predictor case, $z_1 = (c - \rho d)/\sqrt{1-\rho^2}$ and $z_2 = (d - \rho c)/\sqrt{1-\rho^2}$. For ρ = 0, this is the identity transformation and the result is the same as Fig. 1. With $d \to \pm\infty$ in (12), $c \to \rho d \pm \sqrt{1-\rho^2}$, so that $z_1 \to \pm 1$ (Figs. 2c,d); the associated values of z_2 are $d\sqrt{1-\rho^2} \pm \rho$. For small and intermediate values of *d* the bounds are less simple. The same holds for $c \to \pm\infty$ so that $z_2 \to \pm 1$ with $z_1 = c\sqrt{1-\rho^2} \pm \rho$.

Some more insight into Fig. 2 is obtained by noting that the corners of the unit rectangle in Fig 2a transform to the corners of the approximate trapezium in Fig 2c; the (1,1) corner becomes (0.58,0.58), the (1,-1) corner becomes (1.73,-1.73) for ρ = 0.5, and the opposite corners (-1,-1) and (-1,1) follow by mirroring. This means that with the z-ratio pair (0.6, 0.6) both variables stay in the model. So it is not even necessary that the z-ratio exceeds 1 for obtaining a non-zero Type-II ML estimate. For ρ = 0.9, the corners become (0.23, 0.23) and (4.36,-4.36). So, for example, the z-ratio pair (0.3,0.3) gives two non-zero Type-II ML estimates, but the pair (4,-4) yields two zero estimates in Type-II ML so pruning both predictors from the model, despite the fact that for this ρ the chi-square test-statistic of the latter point is about 9 times that of the former. This is a remarkable property of Type-II ML.

Note that the white upper-right and lower-left corners in Figs 2c,d come from the small white wedges in Figs 2a,b in the same corners. In Fig. 2b the wedge is very small: if the correlation among predictors is high, both predictors to stay in the model when *c* and *d* are very close or very different, or both should be very large.

In conclusion, if the (estimated least squares) effect of one predictor is very strong, the bound for the additional correlated predictor comes close to the bound for the uncorrelated case (|z-ratio| > 1 for a predictor to stay in the model). If, by contrast, neither predictor has a large effect, then Type-II ML prunes the one with the smallest effect. If, for positively correlated predictors, they have virtually identical estimated effects, then both predictors stay in the model, even if their z-ratio is

as small as 0.6 and 0.3 for ρ = 0.5 and 0.9, respectively. However, if for positively correlated predictors, the estimated least squares effects are of opposite sign, Type-II ML excludes both predictors, except when the z-ratios are large.

Experiments

In the following we compare the performance of RVM with LASSO and PLS on simulated and real data. Computation was carried out in R (R Development Core Team 2010) using the packages lme4 (Bates et al. 2011), glmnet (Friedman et al. 2010) and pls (Wehrens and Mevik 2006). The kernelized version of RVM was carried out with the function rvm in the kernlab package (Karatzoglou et al. 2004). Results are for two types of kernels: RVM_{rbf} (Gaussian radial basis kernel) and RVM_{lin} (the linear or dot product kernel).A prototype statement to carry out RVM (by Type-II ML) in lme4 with $M = 2$ is

```
lmer(t ~ (0 + x1 | v) + (0 + x2 | v), data=train, REML=FALSE)
```

where t is the target, x1 and x2 predictors, v is an all ones *N*-vector and train is a data frame containing these vectors. The argument REML shows that RVM could also be fitted using Residual Maximum Likelihood (Searle et al. 2008). REML estimates of variance components are generally less biased than ML estimated. In the experiments we show results from both Type-II ML and REML. In RVM context the differences are expected to be small.

Simulation study

We first checked that lmer follows our theoretical analysis that RVM with orthogonal predictors sets the variance of predictors to zero if their |z-ratio| ≤ 1. For this, we generated data sets with R-package mvrnorm with *M*=6 orthogonal predictors and target **t** such that the z-ratio's in a least squares fit were 0.90, 0.94, 0.98, 1.02, 1.06, and 1.10. For large *N* (*e.g. N* = 100 and 1000), lmer followed the theory in all such data sets. For small *N*, the two small differences between our theory and lmer play a role. First we could not fix the error variance to 1 as we did in our theory and, secondly, we could not omit the intercept. The REML- and ML- estimates for the error variance by lmer were biased downward with, as expected, less bias in REML than in ML, and with less bias for larger *N*. For small *N* (we tested with *N* = 8 and 20), our theory still turned out to work for the *estimated* z-ratio, that is, the z-ratio in which the estimated error variance $\hat{\sigma}^2$ is inserted for σ^2. The variance estimates by both REML and ML were in accordance with equation (9) with $v_m = \hat{\sigma}^2/\boldsymbol{\phi}_m^T\boldsymbol{\phi}_m$, except in occasional cases of non-convergence. So, lmer followed the theory for orthogonal predictors also quantitatively.

Next, we simulated data where all predictors were assumed to be independent and Gaussian distributed, but not necessarily orthogonal in each particular data set due to randomness or $N < M$.

Table 1. Median mean-squared prediction errors for the simulated examples with independent predictors for different methods (100 replications). In parentheses are the corresponding standard errors (of the medians) computed via 1000 bootstrap resamples of the 100 mean squared errors. The null model used the mean for prediction.

				Type-II	
M	null model	LASSO	PLS	ML	REML
1	10.33 (0.062)	1.09 (0.011)		1.09 (0.010)	1.09 (0.010)
5	10.31 (0.086)	1.17 (0.024)	1.32 (0.032)	1.14 (0.031)	1.14 (0.027)
10	10.45 (0.087)	1.27 (0.021)	2.04 (0.104)	1.32 (0.035)	1.31 (0.035)
20	10.32 (0.079)	1.45 (0.045)	4.97 (0.275)	2.44 (0.111)	2.38 (0.101)
100	10.39 (0.063)	1.52 (0.045)	8.81 (0.123)	2.21 (0.080)	2.14 (0.053)

We simulated data from the true model $\mathbf{t} = \boldsymbol{\Phi w} + \sigma\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim N(0,1)$, $\sigma = 1$ with $w_1 = 3$ and $w_m = 0$ for $m > 1$ where $m = 1, \ldots, M$, and M = 1, 5, 10, 20 and 100 predictors. The examples thus differ in the number of weights equal to zero (noise predictors). The lme4 implementation of type II ML did not allow much higher M. We set N = 20 to still get a wide range of M/N. For each example, 100 datasets were generated. For computing mean-squared error of prediction of the target (MSEP), each dataset was split into training data of N = 20 observations and test data of 1000 observations and MSEP was calculated from the test data using the weights estimated from the training data.

Table 1 and Fig. 3 summarize the results. Type-II ML and REML behaved similar and almost identically to LASSO for M= 1 and 5 but behaved worse for $M \geq N$. PLS had the worst performance in all examples.

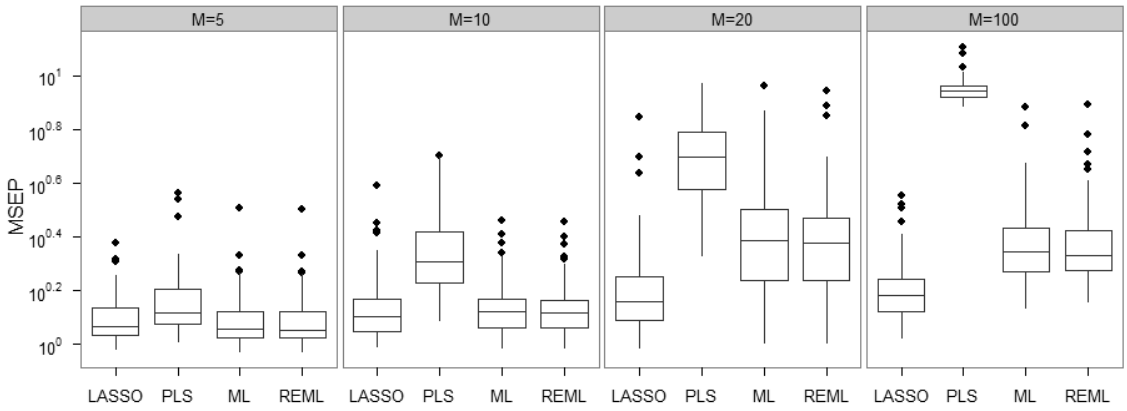

Fig. 3. Box plot of the mean-squared prediction error (MSEP) for LASSO, PLS, ML and REML of the 100 simulations with independent predictors.

The next three examples are similar to those in Zou and Hastie (2005), where simulated 100 data sets are simulated from the model $\mathbf{t} = \boldsymbol{\Phi w} + \sigma\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon}\sim N(0,1)$. These examples are:

In example 1, N = 20, M = 8, $\mathbf{w}$ = (3, 1.5, 0, 0, 2, 0, 0, 0)^T and the predictors are Gaussian with $\mathrm{corr}(\phi_n, \phi_m) = \rho^{|n-m|}$ with ρ= 0.5. We set σ = 3 and this implies SNR ≈ 1.5.

Example 2 is the same as example 1, except that $w_m = 0.85 \ \forall \ m$ (SNR ≈ 1.3).

In example 3, N = 50, M = 40, $w_m = 3$ for m = 1, …, 15 and $w_m = 0$ for m = 16, …, 40 and SNR ≈ 1.7. The first 15 predictors are three equally important groups of 5 predictors each, which are generated as follows:

$\phi_m = h_1 + \epsilon_m^{\phi}$ with $h_1\sim$N(0,1), $m = 1, \ldots, 5$

$\phi_m = h_2 + \epsilon_m^{\phi}$ with $h_2\sim$N(0,1), $m = 5, \ldots, 10$

$\phi_m = h_3 + \epsilon_m^{\phi}$ with $h_3\sim$N(0,1), $m = 10, \ldots, 15$

and $\epsilon_m^{\phi}\sim$N(0,0.16) for $m = 1, \ldots, 15$. In this model, the pairwise correlations within groups are 0.86 and the correlations between groups are 0. The remaining 25 predictors are pure noise features.

The next four examples are from ter Braak (2009) and use a latent variable model. In these examples the target was generated from four independent standard Gaussian latent variables $h_1, \ldots, h_4$ by

$t_n = \sum_{l=1}^{4} \psi_l h_{nl} + \epsilon_n$ with $\epsilon_n\sim$N$(0, \sigma^2)$

and fixed $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_4)^T$, and the predictors were generated as

$\phi_{nk}^{(l)} = \tau_{lk} h_{nl} + \epsilon_n^{\phi}$ with $\epsilon_{nk}^{\phi}\sim$N$(0, 1-\tau_k^2)$ $(n = 1 \ldots, N, l = 1, \ldots, 4, k = 1, \ldots, m_l)$,

and fixed $\{\tau_{lk}\}$, yielding predictors with unit variance. The following four examples differ in the number of predictors per latent variable (m_l), the weights ($\boldsymbol{\psi}$, $\{\tau_{lk}\}$), and the number of noise variables added.

In Example 4, N = 50, M = 75, $\boldsymbol{\psi} = (22.9, 22.9, 22.9, 22.9)^T$ and σ = 15, so that signal to noise ratio (SNR ≡ sd(E($\mathbf{t}$)/σ) is 3. The first latent variable h_1 generate m_1 = 5 predictors with $\tau_{1k} = 0.85 \ \forall \ k$ (SNR= 1.6). The second, third and fourth latent variables generate 10, 20 and 40 predictors in the same way by using q = 2, 4 and 8 repetitions of the τ coefficients, respectively. In this setup, the population least square weights for the predictors associated with the first latent variable are $\mathbf{w}$ = (5, 5, 5, 5, 5)^T and the weights for the predictors associated with the other three latent variables are equal to ~5/q, more precisely 2.59, 1.32 and 0.67. The within-group correlations are 0.72. This example has 75 nonzero coefficients and no zero coefficients.

Example 5 is as example 4, except that 75 nuisance predictors are added.

Example 6 has N = 50 with M =75 predictors and target generated as in example 4 but with different $\{\tau_{lk}\}$, σ = 15 (SNR ≈3.2). For first latent variable $\{\tau_{1k}\} = (20, 20, 0, -20, -20)$. The second latent variable generated two block of five coefficients; in each block $\{\tau_{2k}\} = (10, 10, 0, -10, -10)$. The third latent variable generated four blocks of five predictors; in each block $\{\tau_{3k}\}$ = (2, 2, 0, 2, 2) and the fourth latent variable generated eight blocks of five predictors; in each block $\{\tau_{4k}\}$ = (1, 1, 0, 1, 1). In this setup, contrasts of correlated predictors derived from the first and second latent variable are important for precise prediction (ter Braak, 2009).

Example 7 is as example 6, except that 75 pure noise features are added.

Table 2 summarizes the results based on 100 simulations of the examples. Note that the numbers in Table 2 are σ^2 higher than those in ter Braak (2009). Type II ML and REML perform comparably to LASSO and PLS in examples 1-2, but do poorly in examples 3-7. RVMrbf does better than either Type II ML or REML, except in examples 1 and 2, and better than RVMlin, except in examples 4 and 6. The performance of PLS is the best in all examples, except in examples 1 and 6 where LASSO dominates all.

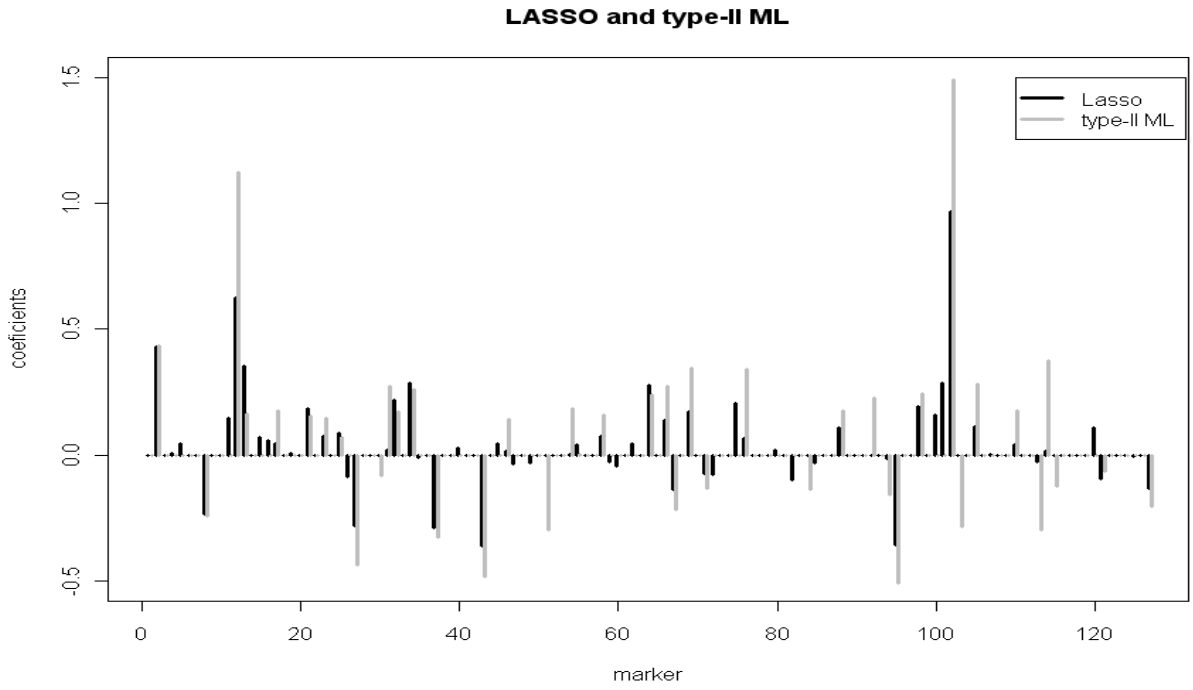

Fig. 4. Marker weights **w** estimated by Type-II ML and LASSO in the barley data.

Real data example

In this example we reconsider the barley dataset from the North American Barley Genome Mapping project to illustrate the performance of Type-II maximum likelihood and LASSO (Xu 2007). The data consists of N = 145 doubled haploid population lines of barley. The target **t** was average kernel weight. The input vector **x** was the genotype of the line, consisting of M = 127 markers. Each marker was coded as $\phi_m(\mathbf{x}) = 1$ for genotype A (TR306 allele), -1 for genotype B (Harrington allele) and 0 for missing genotype. The mean squared error of prediction, estimated using 10-fold cross-validation was 1.62 for Type-II maximum likelihood and 0.68 for LASSO (the LASSO penalty being estimated by an inner loop of cross-validation). Fig. 4 shows the estimated weights of markers for the two methods. Both the methods perform similar in terms of sign and have same direction for coefficients. The LASSO pattern of weights is more shrunken towards zero as compared to Type-II ML. Type-II ML has thus higher peaks.

Discussion

RVM has the attractive property that it automatically selects relevant predictors. Its hyperparameters are estimated by Type-II ML (empirical bayes). By contrast, methods such as LASSO require crossvalidation to set the penalty hyperparameter. We showed analytically that RVM selects predictors on the basis of the least-squares z-ratio ($|z|>1$) in the case of orthogonal predictors and, for M = 2, that this still holds true for correlated predictors when the other z-ratio is large. We also found that RVM prunes the weaker of two highly correlated predictors. In a kernel setting, predictors are likely to be highly correlated, so RVM prunes there. In our simulated and real data, we found that RVM gave higher prediction error than LASSO.

The threshold of 1 for the z-ratio is a kind of minimum that is also implicit in the AIC criterion. For $M > N$, it appears too weak. For example, Donoho (1995) advocated pruning based on $|z| < \sqrt{2\log(M)}$ based on the idea that, for large M, the maximum of M independent standard Gaussian deviates is below this threshold with probability close to 1. More recent work proposes thresholds based on the ratio of the actual and potential model sizes (Abramovich et al. 2005). RVM does not have this property.

In line with the original ideas in Tipping (2001), Xu (2007, 2010) extended the RVM approach by adding a (hyper)prior for the variance components. With a uniform prior for the variances his approach reduces to RVM, whereas it relates to the adaptive sparseness method (Figueiredo 2003) with a Jeffrey's prior. The prior adds a penalty to the marginal likelihood; the penalized marginal likelihood is maximized to obtain the variance components. The prior provides the means for threshold values higher than 1, although we were not yet able to show that analytically.

Table 2 Median mean-squared prediction errors (MSEP) for the simulated examples 1-7 for six methods based on 100 replications. In parentheses are the corresponding standard errors (of the medians) estimated by using 1000 bootstrap resamplings of the 100 MSEPs. For each example the smallest mean-square is in bold (NA = not available as rvm ended with an error).

Method	Example 1		Example 2		Example 3		Example 4		Example 5		Example 6		Example 7	
	MSEP	se	MSEP	se	MSEP	se	MSEP	se	MSEP	se	MSEP	se	MSEP	se
LASSO	**12.4**	(0.34)	13.5	(0.31)	311.0	(4.3)	483.4	(8.8)	436.6	(8.2)	**663.8**	(14.3)	1134.9	(19.3)
PLS	13.4	(0.38)	**11.0**	(0.30)	**273.4**	(4.4)	**351.5**	(4.9)	**361.7**	(6.7)	750.1	(13.8)	**989.7**	(19.9)
Type-II ML	12.9	(0.40)	13.9	(0.35)	380.1	(7.5)	1132.7	(27.0)	601.4	(11.2)	1129.5	(27.1)	1278.4	(37.1)
REML	12.8	(0.37)	13.8	(0.29)	379.7	(8.3)	1155.9	(37.6)	599.5	(10.1)	1110.7	(32.2)	1314.1	(36.1)
RVM_{rbf}	24.0	(0.81)	20.6	(1.12)	347.0	(3.2)	605.7	(7.3)	437.3	(6.17)	1061.0	(11.5)	1044.2	(10.7)
RVM_{lin}	NA		NA		431.5	(7.4)	512.3	(6.4)	603.1	(8.5)	884.2	(12.9)	1279.6	(9.5)
σ^2	9		9		225		225		225		225		225	
N	20		20		50		50		50		50		50	

Penalized methods are often given additional underpinning as giving maximum a posteriori (MAP) estimates in the Bayesian framework (Zou and Hastie 2005). In the same vein, RVM yields variance estimates that are MAP under a uniform prior for the variances $\boldsymbol{\alpha}$. But what happens in terms of precisions? The posterior density would change with a Jacobian term involving $\prod a_m^2$ that accounts for the transformation to precision and therefore the MAP would change when back-transformed to variance. By contrast, penalized methods are invariant under transformation. The Bayesian underpinning of penalized methods is thus rather thin.

This raises the question whether RVM and Xu's extensions can be thought of as approximations to a fully Bayesian model. Xu (2007, 2010) uses independent scaled inverse chi-square distributions as priors for the variances, which is equivalent to gamma distributions for the precisions. The prior for α_m is thus inverse gamma

$$p(\alpha_m|a,b) \propto \alpha_m^{-(a+1)} \exp\left(-\frac{b}{\alpha_m}\right), \quad (13)$$

which is proper for $a>0$ and $b>0$ and leads to t-priors for the weights. For obtaining more shrinkage, Xu (2007, 2010) used improper priors with $b=0$ and $-1 \leq a \leq 0$. The model is equivalent with the improper δ-prior (ter Braak et al. 2005, ter Braak 2006).

$$p(\alpha_m|\delta) \propto \alpha_m^{\delta-1}. \quad (14)$$

Their fully Bayesian treatment showed that the posteriors for $\boldsymbol{\alpha}$ and $\mathbf{w}$ are proper if and only if $0 < \delta \leq 1/2$ or equivalently $-1/2 \leq a < 0$ with $b = 0$ in (13) and that the model gives attractive sigmoidal shrinkage for small δ, similar in form of that of the SCAD penalty (Fan and Li 2001). Note that the uniform prior ($\delta = 1$) for α_m (RVM) and Jeffey's prior are excluded ($\delta = 0$). The uniform for the standard deviation $\alpha_m^{1/2}$ ($\delta = 1/2$) is not excluded, but does not shrink. We conclude that the empirical Bayes approach in RVM and its extensions by Xu (2010) are not supported as approximations to a fully Bayesian approach; the fully fletched Bayesian model does not even exist for the values used for the parameter of *a* and *b*.

Parameters of priors such as (13) and (14) can no longer be estimated in a Bayesian way if they are improper. The reason is that it is impossible to add an additional level to the Bayesian model and to assign them a hyper prior so as to obtain the posterior distribution of $\boldsymbol{\alpha}$ for the assigned hyper prior. It is therefore of interest to define a proper prior for the variances. In terms of the scaled variance ratios $\gamma_m = (\boldsymbol{\phi}_m^T \boldsymbol{\phi}_m)\alpha_m/\sigma^2$ a useful proper prior is

$$p(\gamma_m|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \gamma_m^{a-1}(1+\gamma_m)^{-(a+b)}, \text{ for } a>0 \text{ and } b>0. \quad (15)$$

For $a = b = \delta$ small, (15) gives very similar shrinkage properties as shown in ter Braak (2006) for (14). This prior is closely related to the beta distribution; if $s_m \sim$ Beta(a,b), then $\gamma_m = s_m/(1-s_m)$ follows distribution (15). Conversely, $s_m = \gamma_m/(1+\gamma_m)$ which can be interpreted

as shrinkage coefficient; it relates the shrunken estimate $\widetilde{w}_m$ to the least-squares estimate $\widehat{w}_m$ via $\widetilde{w}_m = s_m \widehat{w}_m$ in the orthogonal predictors case. Whereas (15) implies a proper Beta(a,b) prior for s_m, (14) implies the improper Beta($\delta, -\delta$) prior. The model with the proper prior is a rival for methods in which discrete mixtures of weights (George and McCulloch 1993, Johnstone and Silverman 2004) or variances (Meuwissen et al. 2001) give sparsity and is of interest for further study; see *e.g.* (Polson and Scott 2009). Such models are needed as this paper suggests that Type-II ML in the linear model with individual variance parameters is not the general answer in high dimensional prediction problems.

Appendix A: Derivation of equation (7)

Here we convert the marginal likelihood $L(\boldsymbol{\alpha})$ from a form that uses $N \times N$ matrices to one that uses $M \times M$ matrices. We start with

$$L(\boldsymbol{\alpha}) = (2\pi)^{-N/2}|\mathbf{C}|^{-1/2}\exp\left(-\frac{1}{2}\mathbf{t}^T\mathbf{C}^{-1}\mathbf{t}\right), \quad \text{(A.1)}$$

where $\mathbf{C} = \sigma^2\mathrm{I} + \boldsymbol{\Phi}\mathbf{A}\boldsymbol{\Phi}^T$. The Matrix Determinant Lemma gives (Golub and van Loan 1989, Roweis 1999)

$$|\mathbf{C}| = |\sigma^2\mathrm{I} + \boldsymbol{\Phi}\mathbf{A}\boldsymbol{\Phi}^T| = |\sigma^2\mathrm{I}||\mathbf{A}|\left|\mathbf{A}^{-1} + \boldsymbol{\Phi}^T\boldsymbol{\Phi}\sigma^{-2}\right| = |\sigma^2\mathrm{I} + \boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{A}|. \quad \text{(A.2)}$$

The Matrix Identity Lemma or Woodbury formula gives (Golub and van Loan 1989, Roweis 1999, Bishop 2006)

$$\mathbf{C}^{-1} = (\sigma^2\mathrm{I} + \boldsymbol{\Phi}\mathbf{A}\boldsymbol{\Phi}^T)^{-1} = \sigma^{-2}[I - \boldsymbol{\Phi}(\sigma^2\mathbf{A}^{-1} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}], \quad \text{(A.3)}$$

so that

$$\mathbf{t}^T\mathbf{C}^{-1}\mathbf{t} = \sigma^{-2}(\mathbf{t}^{\mathrm{T}}\mathbf{t} - \mathbf{t}^{\mathrm{T}}\boldsymbol{\Phi}(\sigma^2\mathbf{A}^{-1} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}). \quad \text{(A.4)}$$

On inserting (A.2) and (A.4) in (A.1) and deleting the terms that do not depend on $\boldsymbol{\alpha}$, we obtain

$$L(\boldsymbol{\alpha}) \propto |\mathbf{I} + \sigma^{-2}\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{A}|^{-1/2}\exp\left(\frac{1}{2\sigma^2}\mathbf{t}^T\boldsymbol{\Phi}(\boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{\Phi} + \sigma^2\mathbf{A}^{-1})^{-1}\boldsymbol{\Phi}^{\mathbf{T}}\mathbf{t}\right), \quad \text{(A.5)}$$

which is (7).

Appendix B: Derivation of equation (8)

If $\boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{\Phi} = \mathbf{I}$, (A.5) decomposes as a product of individual likelihoods $L(\alpha_m)$ with

$$L(\alpha_m) \propto (1 + \sigma^{-2}\boldsymbol{\phi}_m^T\boldsymbol{\phi}_m\alpha_m)^{-1/2}\exp\left(\frac{1}{2\sigma^2}\mathbf{t}^{\boldsymbol{T}}\boldsymbol{\phi}_m(\boldsymbol{\phi}_m^T\boldsymbol{\phi}_m + \sigma^2/\alpha_m)^{-\mathbf{1}}\boldsymbol{\phi}_m^T\mathbf{t}\right). \quad \text{(B.1)}$$

With $\widehat{w}_m = \boldsymbol{\phi}_m^T\mathbf{t}/\boldsymbol{\phi}_m^T\boldsymbol{\phi}_m$, the least-squares estimate, and $v_m = \sigma^2/\boldsymbol{\phi}_m^T\boldsymbol{\phi}_m$, the variance of $\widehat{w}_m$, (B.1) can be written as

$$L(\alpha_m) \propto (1 + v_m^{-1}\alpha_m)^{-1/2}\exp\left(\frac{\widehat{w}_m^2 v_m^{-2}\alpha_m}{2(1+v_m^{-1}\alpha_m)}\right), \quad \text{(B.2)}$$

which is (8).

Appendix C: Derivation of equation (10)

Next, we consider the case of two correlated predictor variables with weights with variance parameters α_1 and α_2. On defining $\boldsymbol{\gamma} = \boldsymbol{\alpha}/\sigma^2$, $c = \boldsymbol{\phi}_1^T\mathbf{t}/\sigma$, $d = \boldsymbol{\phi}_2^T\mathbf{t}/\sigma$, (A.5) becomes

$$L(\gamma_1,\gamma_2) \propto \left| I + \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\begin{bmatrix} \gamma_1 & 0 \\ 0 & \gamma_2 \end{bmatrix}\right|^{-1/2} \exp\left[\frac{1}{2}\begin{bmatrix} c \\ d \end{bmatrix}\left[\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} + \begin{bmatrix} \gamma_1 & 0 \\ 0 & \gamma_2 \end{bmatrix}^{-1}\right]^{-1} [c \quad d]\right]. \quad \text{(C.1)}$$

Differentiating $L(\gamma_1,\gamma_2)$ with respect to γ_1 with Mathematica and then setting derivatives to zero gave a ratio for $\hat{\gamma}_1$ with numerator

$$-1 + c^2 - 2\hat{\gamma}_2 + 2c^2\hat{\gamma}_2 - \hat{\gamma}_2^2 + c^2\hat{\gamma}_2^2 - 2cd\rho\hat{\gamma}_2 - 2cd\rho\hat{\gamma}_2^2 + \rho^2\hat{\gamma}_2 + \rho^2\hat{\gamma}_2^2 + d^2\rho^2\hat{\gamma}_2^2 \quad \text{(C.2)}$$

and denominator $(1 + \hat{\gamma}_2 - \hat{\gamma}_2\rho^2)^2$. Simplifying with Mathematica did not help and was done by hand by collecting terms that involved *c* or *d* and those that did not. The terms involving *c* or *d* are

$$c^2\hat{\gamma}_2^2 - 2cd\rho\hat{\gamma}_2^2 + d^2\rho^2\hat{\gamma}_2^2 = \hat{\gamma}_2^2(c - d\rho)^2\,, \quad \text{(C.3)}$$

$$2c^2\hat{\gamma}_2 - 2cd\rho\hat{\gamma}_2 = 2c\hat{\gamma}_2(c - d\rho) \quad \text{(C.4)}$$

and c^2, resulting in

$$\hat{\gamma}_2^2(c - d\rho)^2 + 2c\hat{\gamma}_2(c - d\rho) + c^2 = (\hat{\gamma}_2(c - d\rho) + c)^2 \quad \text{(C.5)}$$

and the terms involving neither *c* nor *d* are

$$-1 - 2\hat{\gamma}_2 - \hat{\gamma}_2^2 + \rho^2\hat{\gamma}_2 + \rho^2\hat{\gamma}_2^2 = -(1 + \hat{\gamma}_2)\big(1 + \hat{\gamma}_2(1 - \rho^2)\big), \quad \text{(C.6)}$$

so that by insertion

$$\hat{\gamma}_1 = \frac{\big(c+\hat{\gamma}_2(c-\rho d)\big)^2 - (1+\hat{\gamma}_2)\big(1+\hat{\gamma}_2(1-\rho^2)\big)}{(1+\hat{\gamma}_2(1-\rho^2))^2}. \quad \text{(C.7)}$$

The expression for $\hat{\gamma}_2$ was obtained by symmetry.

Acknowledgements

We thank Laura Astola for help with Mathematica in the extended RVM model and Luke Tierney for suggesting the Beta prior for the shrinkage coefficient. Jamil's research was supported by a grant from Higher Education Commission of Pakistan through NUFFIC (The Netherlands).

Chapter 6

Trait-environment relationships and tiered forward model selection in linear mixed models

Tahira Jamil, Wout Opdekamp, Ruurd van Diggelen, Cajo J.F. ter Braak

Abstract

Trait-environment relationships are usually complex due to the high number of interacting environmental variables and traits. To understand patterns of variation in species biomass in terms of species traits and environmental variables a one-to-one approach might not be sufficient and a multi-trait multi-environment approach will be necessary.

A multi-trait multi-environment approach is proposed, based on a mixed model for species biomass. In the model, environmental variables are species-dependent random terms, whereas traits are fixed terms and trait-environment relationships are fixed interaction terms. In this approach, identifying the important trait-environment relationship becomes a model selection problem. Because of the mix of fixed and random terms, we propose a novel tiered forward selection approach for this. In the first tier, the random factors are selected, in the second, the fixed effects and in the final tier non-significant terms are removed using a modified Akaike information criterion. We complement this tiered selection with an alternative selection method, namely Type-II maximum likelihood.

A mesocosm experiment on early community assembly in wetlands with three two-level environmental factors is analyzed by the new approach. As the number of traits exceeded the number of species, a backward selection approach was not even possible in this case. The results are compared with those of two existing one-to-one approaches, namely the fourth corner problem and the linear trait-environment method, which use permutation for determining statistical significance of trait-environment relationships. Traits related to germination and seedling establishment are selected as being most important in the community assembly in these wetland mesocosms.

Introduction

Understanding the processes that drive community assembly has been and still is a major challenge in community ecology (Diamond 1975, Weiher and Keddy 1995). Many studies have already shown the importance of environmental factors in controlling and shaping species composition (Weiher and Keddy 1995, Kotowski et al. 2010).

The use of species-traits instead of species identity in community ecology research has many advantages as the latter reflect a species adaptation to its environment (Menezes et al. 2010). Hence, a trait-based approach not only allows a comparison of the same process in different vegetation types (e.g. (Díaz et al. 2001, Lavorel and Garnier 2002, Kahmen and Poschlod 2008) but also gives insight into the mechanisms responsible for such patterns (Kahmen and Poschlod 2004) and allows predictions about possible future changes.

In the last decade plant trait data have become more easily available, especially in Western Europe (LEDA, BIOLFLOR, …). This has further strengthened the growing interest of ecologists to study the responses of plant functional traits to environmental conditions (Weiher et al. 1999, Violle et al. 2007).

Despite this increasing interest, our knowledge of plant community assembly is still hampered as the quantification of the effect of plant traits on community assembly stays a real statistical challenge (Dray and Legendre 2008). Most studies on trait-environment relationship, especially model-based ones, are limited to single species. Knowledge about the effect of traits on plant community assembly stays limited. Empirical evidence is limited as most of the studies are observational and correlative (McGill et al. 2006, Vile et al. 2006). Therefore there is an immense need for randomized multispecies experiments that study trait-environmental links and for statistical methods that can link the experimental environmental factors to traits in such multispecies studies. This paper proposes such methods and applies them to a factorial three-year mesocosm study of plant communities with three environmental factors, each on two levels. The experimental measurements are the biomass per species in each of the three years.

The linkage between the traits and the environment is expressed differently in different statistical models. It is a Pearson correlation in the fourth corner problem (Dray and Legendre 2008) and an interaction term in the mixed model approach (Chapter 2). Environment-trait relationships are usually very complex due to the high number of interacting environmental variables and traits. To understand patterns of variation in species density a multivariate approach will be necessary as a

one-to-one approach might not be sufficient. This way, selecting important trait-environment interactions becomes a model selection problem in mixed models.

Mixed models are extremely flexible and form a computationally attractive tool to model complex and large datasets. Their potential applications in ecology are numerous. The resulting flexibility and model complexity makes model selection even more vital (Greven and Kneib 2010). In mixed models, model selection not only includes selecting the best mean structure but also the most optimal variance-covariance structure (Wolfinger 1993, Wolfinger 1996). Despite the fact that mixed models have been available for a few decades, there is surprisingly little literature available concerning model specification, i.e. "Which set of candidate models should be considered", "How to select a model" and "What is the best model to use?" These are *the* critical questions in making valid inference from data. This also includes the *variable selection problem* in mixed model analysis.

One of the goals of model selection is a trade-off between model complexity and accuracy. Depending on the modeling objective, different procedures to select an optimal model subject to a particular criterion are available. However, it is important to adopt a model selection procedure that reflects the ultimate objective of the modeling process (Hoeting et al. 2006). Model selection is often done though sequential testing either stepup (forward) or stepdown (backward) regression methods (Hosmer and Lemeshow 2000). For simple problems, the outcomes of model selection using these two approaches might happen to be similar; however, in more complex situations, with many candidate models, the results of the two approaches may be quite different. Yet, selecting the best model from all possible models with different fixed and random effect factors is computationally forbidding as the number of models grows exponentially with the number of factors.

This paper develops a novel model selection method called tiered forward selection. The method uses a modified Akaike information criterion. In the first tier, the random factors are selected, in the second, the fixed effects and in the final tier non-significant terms are removed. In our case study, the random factors are the environmental factors while the fixed effects are related to traits and trait-environment interactions. We complement this tiered selection with an alternative one-shot method, namely Type-II maximum likelihood (Type-II ML).

The paper is structured as follows. After a short description of the mesocosm example data and association data screening, the linear mixed model, the tiered forward selection and Type-II ML are presented. Next we describe two simple, existing methods for detecting trait-environment relationships that are not based on mixed models. These methods, the fourth corner method and the linear trait-environment method (LTE), use permutation for determining statistical significance.

After presenting the results we discuss statistical issues and shortly interpret the results in biological terms.

Material & Methods

Data

Data are from an outdoor mesocosm experiment investigating early community assembly from a pool of 28 floodplain species covering a wetness gradient. Experimental variables were i) Canopy presence (with and without canopy) at initial germination, ii) Waterlogging (with and without waterlogging), iii) Mowing (with and without summer mowing) for a full-factorial design (2 x 2 x 2) using 10 replicates. Canopy (C/nC) consisted of Poa pratensis, Lolium perenne and Alopecurus pratensis. Grasses were pre-grown for 6 weeks and cut at 10 cm height at the sowing date (this grass species mixture ensured the persistence of a grass canopy in both dry and wet conditions). In the other mesocosms, soil was kept bare until the sowing date through weeding.

In waterlogged (W/nW) soils water level was maintained at 5cm below soil surface while water in other soils was unobstructed from percolating through the soil profile.

Mowing (M/nM) involved annual mowing of vegetation to 2cm in June–July.

Aboveground biomass of all mesocosms was harvested at the end of August 2006, 2007 and 2008, and was sorted to species level with the exception of Poa, Lolium and Alopecurus (grouped together). Dry mass was measured after 72 hours of drying at 70°C. A more detailed description of the experimental set-up can be found in Kotowski et al (2010).

Traits were either measured (for details see Kotowski et al 2010) or extracted from several plant trait databases: Biobase (van Duuren et al. 2003), BiolFlor (Kühn et al. 2004), CloPla (Klimešová and Klimeš 2006), LEDA (Kleyer et al. 2008) and SID (Liu et al. 2008). A description of the different traits can be found in Table 1.

Data screening

Prior to analysis, trait data was screened for zero-variance predictors and for multicollinearity among predictors. Predictors with a single unique value (also known as “zerovariance predictors”) and near zero-variance predictors (for details see (Kuhn 2008)) can cause numerical problems and lead to misinterpretation. Both near zero-variance and zero-variance predictors were removed from the dataset. Predictor detection was performed using the `caret` package for R (Kuhn 2011). Multicollinearity among predictors is a problem in mixed models. From each pair of trait predictors with correlation greater than 0.80, one predictor was removed using the “findCorrelation” function in the `caret` package in R (Kuhn 2011). This function removes the predictor that has highest mean pairwise correlation with the other predictors.

Table 1. Traits used for analysis with code and description

	Germination traits	
Z1	SW	Seed weight (mg)
Z2	GP	Total germination percentage in full light (%)
Z3	T50	Time of 50% germination in light (days)
Z4	WGR	Wet germination ratio (germination in wet mesocosms/germination in dry mesocosms)
Z5	DGR	Dark germination ratio (dark germination/light germination)
	Seeding traits	
Z6	H7	Average height of seedlings at 7th day from germination (mm)
Z7	LWR7	Mean leaf weight ratio at 7th day (leaves and cotyledons)
Z8	LAR7	Mean leaf area ratio, i.e. mean quotient of the total leaf area per plant and the total weight per plant at 7th day (mm^2 mg^{-1})
Z9	AGR	Mean actual growth rate of seedlings between the 7th and 22nd day (mg day^{-1})
Z10	RGR	Mean relative growth rate of seedlings between the 7th and 22nd day
Z11	LWR	Mean leaf weight ratio, increase between the 7th and 22nd day
Z12	LA7	Mean leaf area of seedlings at 7th day (mm^2)
	Adult traits	
Z13	CH	Canopy height (maximum value) (m)
Z14	LDMC	Leaf Dry Matter Content (mg g^{-1})
Z15	SLA	Specific Leaf Area (adult plants only) (mm^2 mg^{-1})
Z16	HEM	Leaf Distribution along stem (Hemi-rosettes)
Z17	STA	Flowering start (1: January-April; 2: May-June; 3: July-September)
Z18	DUR	Flowering duration (1: Short (1-2 months); 2: Medium (3-4 months); 3: Long (>4 months))
Z19	SEX	Reproductive type (sexual)
Z20	AB0	BudBank Vertical Distribution - Aboveground - 0 (0 buds)
Z21	AB2	BudBank Vertical Distribution - Aboveground - 2 (>10 buds)
Z22	GR2	BudBank Vertical Distribution - Groundlevel - 2 (>10 buds)
Z23	BE0	BudBank Vertical Distribution - Belowground - 0 (0 buds)
Z24	BE1	BudBank Vertical Distribution - Belowground - 1 (1-10 buds)
Z25	GRS	BudBank Seasonality - Groundlevel - seasonal
Z26	BES	BudBank Seasonality - Belowground - seasonal
Z27	LAT2	Lateral Spread (2: 0.01-0.25m year-1)
Z28	LAT3	Lateral Spread (3: >0.25m year-1)

Trait predictors were checked for normality by making histograms. Predictors departing from normality (WGR, H7, AGR, and LA7) were log-transformed (Table 1). Once the final set of predictors was determined, predictors were centered and scaled using their mean and standard deviations (Kuhn 2008). Species with small average biomass (< 0.1) were removed from the analysis and data set is 23 species. Species biomass was log-transformed (log(y+0.01)).

The three experimental factors are indicated by c, w and m and the levels of each are coded numerically as -1 (nC/nW/nM) and 1 (C/W/M) without loss of generality. In this coding, c, w and m are orthogonal and also orthogonal to the interactions cw, cm and wm.

Linear mixed models for trait-environment relations

Single trait-environment relationships

The model for y_{ij}, the species biomass for the j^{th} species in the ith site is

$$y_{ij} = \alpha_j + \beta_j X_i + \gamma_i^{site} + \varepsilon_{ij}, \qquad i = 1,\dots,n,\ \ j = 1,\dots,m \tag{1}$$

where X_i is a known environmental factor, γ_i^{site} is the site effect and ε_{ij} is the error term. This model is of interest for our case study as each of the factors in the mesocosm experiment was coded as if it were quantitative, with the levels of the factors coded as -1 and +1. See Chapter 2 in case X is a multilevel factor. We assume that the intercept α_j and slope β_j for the j^{th} species depend on the known value Z_j of a particular trait (Jamil et al 2011)

$$\begin{pmatrix}\alpha_j \\ \beta_j\end{pmatrix} \sim N\left(\begin{pmatrix}a_0 + a_1 Z_j \\ b_0 + b_1 Z_j\end{pmatrix}, \begin{pmatrix}\sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2\end{pmatrix}\right)$$

and also assume $\gamma_i^{site} \sim N\left(0, \sigma_\gamma^2\right)$ and $\varepsilon_{ij} \sim N(0, \sigma^2)$. This is a random intercept and random slope model where trait is a predictor for the random intercept and slope. Inserting α_j, β_j and γ_i^{site} in equation 1 gives:

$$y_{ij} = a_0 + a_1 Z_j + b_0 X_i + b_1 Z_j X_i + \gamma_i^{site} + \varepsilon_j^{\alpha} + \varepsilon_j^{\beta} X_i + \varepsilon_{ij}, \tag{2}$$

where b_1 represents the trait-environment interaction. The above model is fitted for each combination of trait and environmental variable. To test the trait-environment interaction (with null-hypothesis: $b_1 = 0$), we fit the model without this term

$$y_{ij} = a_0 + a_1 Z_j + b_0 X_i + \gamma_i^{site} + \varepsilon_j^{\alpha} + \varepsilon_j^{\beta} X_i + \varepsilon_{ij} \tag{3}$$

and then compare the two models by an analysis of variance resulting in a P-value for the likelihood ratio (LR) test of model with trait-environment interaction term against the null model without this term.

Multiple trait-environment relationship

In community assembly traits and environmental variables should not be considered in isolation as they influence and often coordinate each other. The development of a multivariate framework, in which multiple traits can be linked to multiple environmental variables is needed (Poff et al. 2006). In this section, we show how to link environmental variables and species traits in a multivariate framework. We use a multi-trait and multi-environment version of the mixed model to select the species traits, environments and trait-environment interactions that significantly contribute to the

species biomass distribution model. Equation 1 for a one-one model can readily be extended to cover multi-trait and multi-environmental variables

$$y_{ij} = a_0 + a_1 Z_{1j} + \cdots + a_k Z_{kj} + b_0 X_{1i} + b_1 Z_{1j} X_{1i} + \cdots + b_k Z_{kj} X_{1i} + \cdots + \varepsilon_j^{\alpha} + \varepsilon_j^{\beta 1} X_{i1} + \cdots + \varepsilon_j^{\beta p} X_{ip} + \gamma_i^{site} + \varepsilon_{ij} \,, \quad (4)$$

where Greek letters are used for random terms. Environmental variables enter the model both as fixed terms and as random terms, and traits enter as fixed terms. In this formulation, the selection of the best traits and environmental factors is a model selection problem

Tiered forward models Selection

When the number of fixed effects and random effects is large, it is computationally very expensive and time-consuming to compute all possible candidate models (Fernandez 2007) due to the presence of random terms and variance-covariance structure (Hoeting et al. 2006, Littell et al. 2006). Furthermore the number of candidate models increases exponentially with increased number of fixed effects and random effects (Yuan and Lin 2005). Different protocols for model selection have been developed, in particular step-up (forward) (West et al. 2006) and top-down (backward) protocols (Diggle et al. 2002). Most stepwise functions take a start model and according to some criteria iteratively add or delete a predictor at each step, to get to the best parsimonious regression model. Backward selection starts from the model with all possible terms included and is only feasible if that model can be fitted. In our case, the number of environmental variables should be less than the number of sites and the number of traits should be less than the number of species. In our data, the former holds true, but the latter does not. Here we develop an approach called tiered forward selection. The analysis was done for the three years separately and for the combined data of all three years.

Model Selection Criteria

Different Information criteria, such as AIC (Akaike 1973), AICC (Hurvich and Tsai 1989), CAIC (Bozdogan 1987) and BIC (Gideon 1978), can be used for model selection in linear mixed model (Gurka 2006). Generally, these information criteria are a function of the likelihood for a given model and a penalty term based on the number of parameters in the model. The general form of information criterion (IC) is

IC = -2 logLL + Penalty factor

where log LL is the loglikelihood derived from fitting the mixed model to the data using either ML or REML.

The use of these criteria is somewhat arbitrary and no formal inference can be made based on these values. Comparison of the values of the criteria for a set of candidate models simply indicates if a superior model exist within the given candidate models (Gurka 2006). For an extensive review and discussion on the theoretical aspects of model selection criteria and procedures see Burnham and Anderson (2002) and Hoeting et al. (2006). The most widely used information criteria is

$AIC = -2\ logLL + 2\times par_n$.

We use the variant SigAIC defined as

$SigAIC = -2\ logLL + 3.84\times par_n$

where par_n is the number of parameter estimates. The variant, SigAIC, which multiplies df by $\chi^2_{1(0.05)} = 3.84$ instead of by 2 (Broman and Speed 2002), guarantees that the addition of a single parameter to a model will result in a lower SigAIC value if and only if that parameter is significant at the 5% level as judged by the LR test.

Phase Tier I: Selection of environmental variance components

The start or null model is the model with crossed random effects for species and sites. In the R package lme4 (Bates et al. 2011), it can be represented as

```
start.model <- lmer( y ~1+(1| sp)+(1|site), data)
```

where y represents the vectorized response data while sp and site indicate species and sites, respectively. REML is the default estimation method. In each consecutive step, the environmental predictor for which the species-dependent random effect term increases the log-likelihood most is added. This means all models with one extra term have to be fitted each step and the best term is retained for the next step. For example for "w"

```
lmer( y0 ~1+(1+w|sp)+(1|site)  ,data)
```

In lme4, such model can be fitted as an update of the start model with the statement

```
update(start.model, .~. +(1+ w|sp) +(1|site)-(1|sp) , data).
```

To generate all models needed in this tier, we used the statement

```
update(start.model,as.formula(paste(". ~ .+(1+",block[j],"|sp)+(1|site)-
(1|sp)"))).
```

Here block is a vector of the candidate predictors, *i.e.* the environmental factors. All models, after fitting the predictors in the block, are arranged in order of the predictive criterion. The best candidate model is compared with the null model. If the best candidate model is statistically significant, it becomes the null model for the next step. This process is repeated until the increase in log-likelihood is no longer statistically significant as judged by LR test.

After this tier, the selected predictors are added as fixed effect. Thus, environmental variables in the model are now the component of both fixed terms and random terms.

Tier II: Selection of Fixed trait effects

In our case study, the start model for the next phase has a specification,

```
start.model<-lmer(y~c+w+cw+(1+c+w+cw|sp)+(1|box),method= "ML", data).
```

Now traits and trait-environment interactions can be added as fixed effects to the model. It is important to note that REML is the default estimation method for mixed models. Generally REML estimates of variance components are preferred to ML estimates. However, in REML it is not legitimate to compare models with different sets of fixed effects as the contrast used to develop the restricted maximum likelihood depends on the fixed effect design matrix (Verbeke and Molenberghs 2000). Therefore the ML estimation method is used in this tier.

Given a starting model and a set (here block) of variables to evaluate, the starting model is updated by adding every single trait variable and trait-environment interaction. To evaluate the importance of a trait the current model is updated with the statement

```
update(start.model, . ~ .+  Z + c:Z )
```

A generic way to update the current model with any single trait is

```
update(start.model,as.formula(paste(". ~ .+ ",block1[j],"+",block2[j]))).
```

Here block1 is for trait main effects and block2 for the trait-environment interaction. Models fitted this way are then ordered based on the chosen predictive criterion, here SigAIC, after which the best fitting model is retained for the next step. The procedure continues until addition of new traits and trait-environment interactions does not significantly improve the model.

Next, the same procedure is repeated for the three-way interaction, but keeping the marginal effect of a trait variable and its two-way interaction,

```
update(start.model,as.formula(paste(". ~ . + ", block1[j],"+", block2[j],
"+", block3[j],"+",block4[j]))).
```

Block1 is for trait main effects, block2 and block3 for two-way interactions and block4 for three-way interaction. This structure ensures the marginality principle (Nelder 2008) which entails a model can include an interaction term (high order term) only if it includes the main effects (and all lower order terms) that compose the interaction. The condition requires that a mixed model with an interaction, say X:Z, must also include the main effects X and Z. In general, we neither test nor interpret the main effects of explanatory variables that are also included in an interaction. The procedure continues until addition of new predictors does not significantly improve the model.

Tier III: Removal of non-significant interaction terms

In this tier non-significant interaction terms are sequentially removed. An example statement in this tier is

```
update(start.model, as.formula(paste(". ~ . - ", block1[j]))).
```

The final model was obtained by sequential removal of non-significant interaction terms. Now we refit the final model using REML estimation, this is needed to obtain unbiased estimates of the covariance parameters. As the ML estimation leads to biased covariance parameter estimates. This can result in smaller estimated standard errors for the estimates of fixed effects in the model. The REML estimate for the fixed effects is not identical to its ML version and differs more from ML as the number of fixed effects in the model increases.

Model selection by Type-II Maximum Likelihood

When a system is described by a statistical model, model complexity leads to a very large computing time and poor estimation, especially if the number of predictors is large relative to data size. As an alternative to and improvement over stepwise methods, shrinkage methods have been proposed. One of these is the Relevance vector machine (RVM) which has gained popularity within the Bayesian framework (Tipping 2001, Bishop 2006, Xu 2007). RVM introduces a Gaussian prior to the regression coefficients, with one variance component for each predictor. The variance component or hyperparameter controls the degree of sparseness. These parameters are commonly adjusted by crossvalidation. In Bayesian framework, the hyperparameters are estimated by using empirical Bayes or, equivalently, Type-II maximum likelihood (Berger 1985).

The Type-II maximum likelihood shrinks the coefficients to zero and readily sets some of the coefficients to zero, namely if their variance component is estimated to be zero (Chapter 5). These zero variance coefficients are equivalent to pruning the corresponding predictors from the model. Hence this method readily helps in pruning predictors from the model and does variable selection and model estimation. A prototype statement in lme4 in linear regression is

lmer(y ~ (0 + x1 | v) + (0 + x2 | v), data, REML=FALSE)

where y is the response and x1 and x2 predictors, v is an all ones *N*-vector and train is a data frame containing these vectors. In our multi-trait – multi-environment context a prototype statement is

lmer(y ~ (0 + z1:x1 | v)+(0 + z1: x2|v)+(0 + z2 : x1 | v)+(0 +z2 : x2|v) +(1+x1+x2|sp)+ (1|site), data, REML=FALSE)

where x1 and x2 indicate two environmental factors, z1 and z2 two traits and z1: x1 and related term indicate the vector that is the product of the corresponding trait and environmental factor.

Fourth Corner Method

The fourth-corner method developed by Legendre et al (1997) and extended in Dray and Legendre (2008) is the oldest integrated trait-environment method. The species data in this method must be non-negative and can thus be presence-absence data, abundance data or species biomass. The fourth-corner statistic measures the link between trait and environment via the species data table with a weighted Pearson correlation coefficient between trait and environmental variable, each vectorized as in the mixed model approach (Chapter 2). The weights are the species data $\{y_{ij}\}$ (presence-absence, abundance or biomass). The role of the species data is thus rather different from that in mixed models and in LTE of the next subsection. In particular, absences or zeroes therefore do not carry any information in the analysis. The method works well for data stemming from unimodal response models (Dray and Legendre 2008). The significance of the trait-environment relationship is tested by a permutation test. Dray and Legendre (2008) offer different permutation scenarios. We used the combined approach implemented in the combine.4thcorner function in the ade4 package in R. It combines the P-values of two fourth-corner models, viz Model 2 (site permutation) and Model 4 (species permutation), as proposed by Dray and Legendre (2008) by taking their maximum (Chessel et al. 2011). This method controls the type I error (Cormont et al. 2011).

The linear trait-environment method (LTE)

The linear trait-environment (LTE) method (Cormont et al. 2011) was developed as an alternative to the fourth corner method to account for negative species data values. The method is linear and as such is more closely related to the liner mixed model than the fourth corner method. However, just like the fourth corner method, LTE has no variance components. The statistical analysis thus proceeds by advanced permutation testing.

LTE starts with a two-step analysis. In the first step regressions per species biomass to each environmental variable give a species specific regression coefficient

$$y_{ik} = a_k + b_k X_i + \varepsilon_{ik}$$

where a_k and b_k are intercept and slope of k^{th} species. In the second step, these regression coefficients are correlated to each trait

$$b_k = c + dZ_k + \delta_k$$

where c and d are the intercept and slope for trait Z and δ_k is a species specific error term with mean 0. LTE integrates both steps in a single model. LTE achieves this integration based on a linear model with main effects for the trait and environmental variable and their interaction. The interaction between a trait and an environmental variable in this model captures the trait-

environment relationship, in particular the trait-dependent effect of environment on species biomass. The significance of this interaction is tested by a permutation test with the same permutation strategy as in the fourth corner problem (Cormont et al. 2011).

Summary of data analysis strategy

We applied the different methods that link functional traits to the experimental environmental factors to data on 23 floodplain species in a factorial mesocosm experiment with three two-level factors (c, w and m for canopy, waterlogging and mowing). We estimated the one-to-one interactions for each trait-environment combination by mixed models, the fourth-corner method (Legendre et al. 1997, Dray and Legendre 2008) and the linear trait-environment method (Cormont et al. 2011). Then we applied the tiered forward selection method to select the traits and environmental variables that significantly contribute to the explanation of species biomass. The environmental variables were c, w, m and their first order interactions. Furthermore, we performed Type-II maximum likelihood analysis (Type-II ML) to select trait-environment interactions. The analysis has been carried out on the three-year mean log-transformed biomass, unless stated explicitly otherwise.

Results

In this study, we explore different methods that link the species traits to the environmental factors in the mesocosm experiment. Table 2 summarizes the results of the one-to-one analyses by mixed models, the fourth-corner method and the linear trait-environment method, and the multi-trait to multi-environment analyses by mixed models and Type-II maximum likelihood. Only significant trait-environment interactions obtained from these analyses are reported together with the sign of the relationship. The factor mowing is excluded from the Table 2 as its variance component in the mixed model approach was very small, smaller than that of the interaction "cw" between the other two factors, canopy "c" and waterlogging "w". Mowing thus has a similar effect on all species. The sign of coefficient for trait-environment interactions are almost consistent between different methods.

From Table 2, H7 and SW appeared important traits that are consistently significant with canopy or/and waterlogging for all methods. Other important traits appear DGR and AGR, although they are less consistent across methods. STA and GRS are significant with waterlogging in except the multivariate mixed model. Type-II ML gave many more trait-environment interactions than the other methods (Table 2 and 4). The coefficients estimated by Type-II ML are plotted in Fig. 1.

Table 3 summarizes the results of the tiered forward model selection in mixed models using SigAIC. The analysis was done for each year and by combining the data for all three years. All

Table 2. Results of analysis from different methods using species biomass, species traits and environmental variables. The sign (+/-) represents the positive or negative significant Pearson correlation/ coefficients between the environmental variable and the species trait. For ease of interpretation, only significant relationships are shown.

	One-to-One Mixed Model			Linear Trait-environment			Fourth- Corner Method			Multivariate Mixed Model			Type-II ML		
	c	w	cw	c	w	cw	c	w	cw	c	w	cw	c	w	cw
SW		−			−	+		−	+	+	−		+	−	+
T50								+	−				−		
WGR		+						+	−		+			+	
DGR	−	−			−	+							−		+
H7	−	−	+	−	−	+	−	−	+	−				−	
LWR7		+									+		−	+	
AGR	−			−	−	+			+	−			−		
LA7		−												−	+
SLA											−				
HEM							+						+		
STA		+						+	−						
DUR		−	+										−	−	
BE0	+						+				+		+		+
BE1														−	
GRS		+	−		+			+	−					+	−
BES			−												

Table 3. Results of the multi-trait and multi-environment analysis using tiered forward model selection in mixed models. The following treatment c: canopy, w: water logging and cw: interaction c×w were always in the model as a fixed effect and random effects.

	year1	year2	year3	all
(Intercept)	-2.66	-2.24	-1.67	-2.20
c	-3.28	-1.97	-1.27	-2.17
w	0.45^{NS}	-0.06^{NS}	0.28^{NS}	0.22
cw	-0.78	-1.02	-0.45^{NS}	-0.75
SW		0.58	0.34	0.11
c:SW				0.86
w:SW		-0.59	-1.24	-0.72
WGR			-0.11^{NS}	-0.12
w:WGR			0.88	0.46
H7	0.83	0.56	0.63	0.77
c:H7	-1.53	-1.49	-1.16	-1.47
w:H7	-0.75	-0.90		
cw:H7	0.60	0.63		
LWR7		-0.51	0.08^{NS}	-0.28
w:LWR7		0.66	1.09	0.67
AGR	0.67			0.47
c:AGR	-1.26			-1.38
SLA			-0.45	-0.13
w:SLA			-0.44	-0.29
HEM			0.36	
BE0		-0.26^{NS}	0.38	0.08
w:BE0			0.87	0.56
c:BE0		0.99		

NS- is for non-significant

Table 4. Trait-environment interactions from Type-II that were not common to other methods

	Type-II ML		
	c	**w**	**cw**
GP			+
RGR	−		
LWR			−
CH			+
LDMC	−		+
AB0		+	
AB2		+	
GR2		+	+
LAT2	−		
LAT3	+	−	

significant traits, except BE0, are germination and seedling establishment traits. Canopy capture more interactions with traits in first and second year of germination and establishment as compared to waterlogging.

Discussion

Statistical

We used three methods to relate each single trait to each single environmental factor. Two of the methods use an explicit linear model for the data (the mixed model and LTE) whereas the third method uses the data are weights. Also, two of the methods use statistical significance testing by a Monte Carlo permutation strategy (LTE and the fourth corner method) whereas the mixed model uses LR testing. What effects do these theoretical similarities and dissimilarities have on the results? The numbers of one-to-one relationships found (Table 2) were 17 in the mixed model, 11 in LTE and 16 in the fourth corner method. The mixed model disagreed in its significance judgement in 14 cases (out of the 48 shown in Table 2) with LTE and in 15 cases with the fourth corner method. LTE and the fourth corner disagreed in 13 cases. The methods thus are about equally dissimilar among one another. The multi-trait multi-environment mixed model yielded 8 significant interactions, so fewer than the one-to-one methods. This is to be expected if one trait can replace another because of their mutual correlation. In our data, traits show some, but no high correlation. Apparently the other aspect of a multivariate model, namely that is can reduce the error variance and thereby can find more significant effects, is less important.

Type-II ML identified many of the trait-environment interactions which were also identified by other methods but also many more. As shown in Chapter 5, Type-II ML tends to be very tolerant for predictors (here interactions) to stay in the model.

Biological interpretation

A major objective of this study was to compare the effects of competition from canopy and waterlogging on assembly processes in a floodplain and how plant functional traits are related to the successful establishment of species. Both wetness and light affected species germination in our experiment, but only few species were directly eliminated at this stage (one under oxic and three under dark conditions). Clearly, canopy presence was a much stronger filtering factor. Especially in the first year it almost totally disabled establishment, which is probably due to high light attenuation. However, more severe root competition could possibly also play a role. Waterlogging is a major constraint on growth and establishment of plant species in wetlands (Lenssen et al. 2003). Only species that are adapted to this environment can occur and thrive. It is relevant to understand which specific combination of traits makes the germination and establishment of plants

resilient to waterlogging and which species traits showed the greatest tolerance to waterlogged conditions. The knowledge which traits determine species seedling and germination might help for conservation planning (Dolédec et al. 1999).

Species able to establish successfully within the grass canopy showed high seed weight, combined with a small individual size and a relatively low actual growth rate. These traits are related to a stress-tolerant strategy, allowing plants to minimize resource requirements and survive in suboptimal conditions. Apparently even the largest seedlings had difficulties in reaching layers with sufficient light. This confirms that large seed size may contribute to seedling establishment in shade through various mechanisms (Westoby et al. 1992). Large seeds with a large nutrient stock are often thought to be advantageous in dense canopies as seedlings should possess enough resources to reach layers with higher light availability (Grime 1979). However, a significant fraction of resources in a large seed may remain in storage instead of being used for immediate seedling development (Garwood 1996) (Green and Juniper 2004), a strategy characteristic for stress-tolerant species. When germinating on bare ground, seedlings of different species compete with each other for light. Yet, the intensity of this competition is much higher in the non-waterlogged treatment. Traits responsible for rapid establishment and outcompeting neighbours appear more important here than those responsible for shade tolerance (Keddy et al. 1994, Stockey and Hunt 1994). A combination of fast growth and large-sized seedlings are prerequisites for success under dry conditions without imposed light stress. In waterlogged soils, specific leaf area (SLA) decreased as waterlogging induces an increased allocation to roots (Lenssen et al. 2003). The ability to germinate in wet conditions is a main determinant of community assembly. This is in accordance with the habitat filter theory which states that the number of species in the local species pool is reduced by habitat constraints.

Hence, all traits (except one) that were selected by the tiered forward model selection describe germination and seedling establishment. This stresses the importance of both two stages as major bottle necks for species recruitment (Grubb 1977, Shipley et al. 1989) and how they may largely determine patterns of biological diversity (Grime 1979, Henry et al. 2004). Moreover, because of their small stature, seedlings can be subject to a totally different light regime and soil resource availability than adult plants, even in the same site.

In conclusion, we have demonstrated different methods that link environmental factors (e.g waterlogging and canopy) to species traits during early assembly process in a wetland mesocosm. Our results clearly stress how the choice of a particular statistical method to analyse the trait-environment link will have consequences for the ecological interpretation of this link. Of the studied methods, the multi-trait multi-environmental mixed model is clearly best suited for predictive usage.

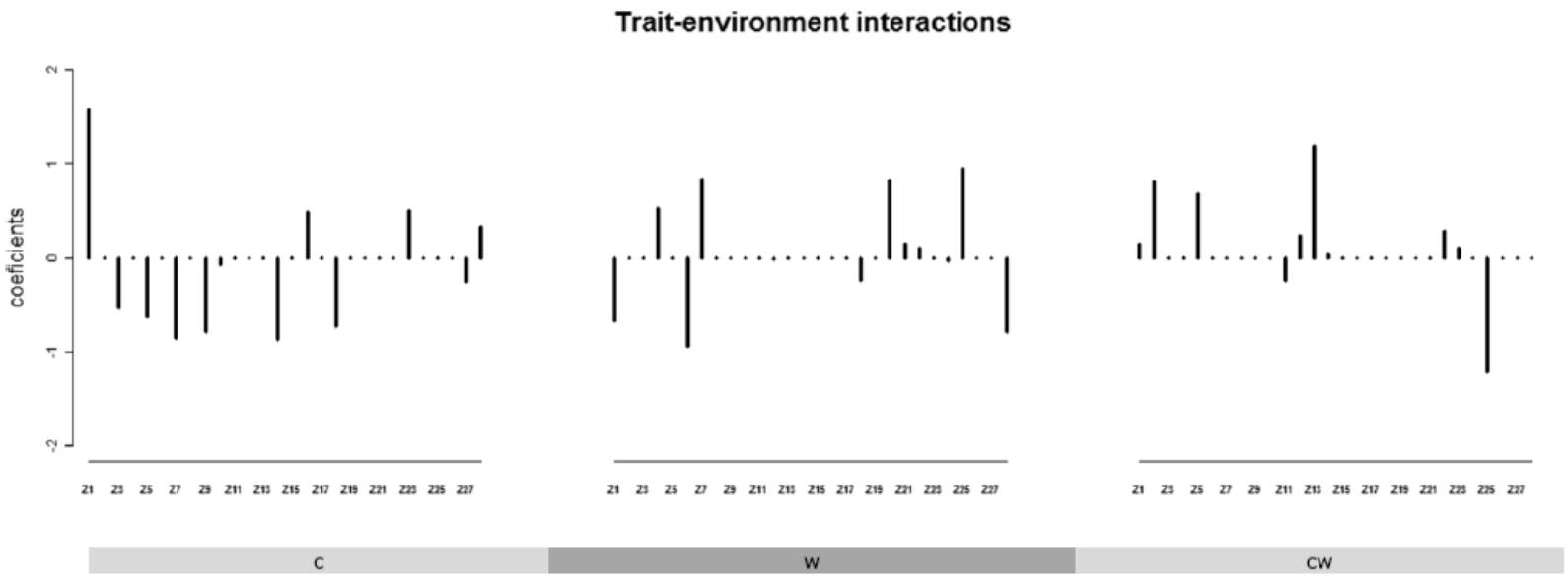

Fig. 1. Estimated effects of trait-environment interaction using Type-II ML.

Chapter 7

General discussion

Species can be characterized by a large number of quantitative and qualitative traits. Following pivotal works by Southwood (1977) and Townsend and Hildrew (1994), trait-based approaches have been increasingly applied to explain and predict response of species to environmental conditions (Weithoff 2003, Follows et al. 2007, Litchman and Klausmeier 2008). These studies cluster the species based on their functional trait and then summarize their response to environmental change. Results from these studies reveal that traits could offer new insights into ecology. In this thesis different methods of modelling the species distribution were approached from a hierarchical modelling perspective. The central theme was to develop models for species distribution that integrate trait-environment relationships.

Chapter 2 of this thesis developed a statistical approach to relate species traits to environment using an extension of the generalized linear model (GLM), namely the generalized linear mixed model (GLMM). GLMMs form a very powerful class of statistical models in ecology and elsewhere (Gelman and Hill 2007, Bolker et al. 2009, Zuur et al. 2009) and are introduced here for modeling and explaining species response along environmental gradients by species traits. Ecological data are often binary (e.g. presence or absence of a species in a site), or counts and have short environmental gradient. Such data sets are most common in practice. GLMMs are the best tool for analysing nonnormal data and that involves random effects and have short environmental gradient. GLMM is based on a sound statistical model that allows, as a standard by-product, questions to be answered about which traits and environmental variables are significantly related and which best explain the species response in a parsimonious model.

Our GLMM can be considered as integrating two steps in to one. The basis of our GLMM is the random intercept and random slope model (step 1). We made the intercepts and slopes dependent on the species traits (step 2). The result is a GLMM which combines steps 1 and 2; it has main effects for traits and environmental variables as well as interaction effects between them and random effects for sites, species and environmental variables. In our GLMM model, the trait-environment relationship is an interaction term and not a correlation. It can be tested for statistical significance using standard software. The GLMM utilizes species trait data efficiently and overcomes the problem of pseudo-replication (Paterson and Lello 2003).

GLMM has several advantages over existing two step approaches. It models directly the variable of interest (occurrence probability, expected abundance), it automatically weighs the different kinds of information for an optimal model fit and statistical significance testing and it provides consistent estimates of the between-species variance of (slope) parameters, without introducing unnecessary random variation by replacing the (slope) parameters by their estimates as in the two step approach and it can be applied with small sample size.

The GLMM regression coefficients for each species are a compromise between the coefficients from a per-species fit and the population average and are also called shrinkage estimate (Pinheiro and Bates 2000). Species which have few presences lead to abnormally high estimates in the GLM fit. The pooling across species in the GLMM estimation gives a certain amount of robustness to species with few occurrences in the data.

Niche theory predicts that species occurrence and abundance show non-linear, unimodal relationships with respect to environmental gradients (Økland 1986, Austin 1987, Minchin 1989, Palmer and Dixon 1990, Begon et al. 1990). Many studies fail to test for unimodal response (Austin 2007). Thus linear relationships are often fitted without justification. Chapter 4 developed explicit testing of unimodality in species response along an environmental gradient, without fitting a unimodal model. In this Chapter we took a simple approach and studied the suitability of GLMM for detecting the unimodality of species response along an environmental gradient and suggested a graphical tool and a statistical test for testing unimodality. There is an indication for unimodality when site effects show a quadratic relationship with the environmental gradient. The test can make even stronger by adjusting the relationship with the site total (the number of species in a site). As an alternative, we could explicitly add the square of the environmental variable as a fixed effect term to the GLMM and judge the significance of its addition. This approach is presumably even more powerful, but necessitates the fit of an extra model. The theory developed in this chapter gives insight in why an ordination method such as correspondence analysis can be viewed either as linear or as a unimodal and why the linear-trait-environment method (LTE) might be applicable even when the data are unimodal.

Each species has preferred environmental conditions in which it can survive and reproduce optimally. Thus it is presumed to occur in a characteristic, limited range of the multi-dimensional habitat space, called its ecological niche, and within this niche, each species tends to be most abundant around a specific environmental optimum (Green 1971). Therefore, the distribution of species along any environmental gradient is usually unimodal, with the maximum at some ecological optimum. The symmetric and bell-shaped unimodal (non-negative) species response curve is a Gaussian response curve with three interpretable parameters: the optimum, height of the response and tolerance or width of the curve (Jongman et al. 1995, Oksanen and Minchin 2002).

Generally, the bell-shaped response curve is not estimated by nonlinear regression, but is estimated by the equivalent (or almost equivalent) polynomial model (ter Braak and Looman 1986, Oksanen and Minchin 2002), which can be easily fitted using generalized linear model and that gives results close to the real Gaussian curve. GLM can model presence-absence, counts or biomass data with an appropriate link function.

It is difficult to fit the Gaussian logistic model (ter Braak and Looman 1986) with linear trait submodels for the parameters with the available (generalized) nonlinear mixed model software. Instead, a Bayesian approach is applied and fitted using OpenBUGS (Bayesian inference Using Gibbs Sampling). Chapter 3 adopted a fully Bayesian approach to model unimodal species response model relating traits to environment. A GLMM is, of course, a linear model. GLMM uses the environmental variables linearly (Chapter 2). When data come from an ecosystem with niche structure, i.e. from a unimodal system, adding polynomial terms as random component to the linear model is less attractive as it leads to coefficients that lack a clear ecological interpretation and have no intuitive meaning (Chapter 3). Moreover, the meaning of the parameter of the linear term depends on the value of that of the squared term and also on the scale used for the environmental variable. It appears therefore rather useless ecologically to make these parameters dependent on the species traits. By contrast, the optimum, the tolerance and the maximum are interpretable parameters and were modeled in terms of the species traits (Chapter 3).

Crucial to our aim is the identification of those traits responsible for explaining the variation in response curve parameters (optimum, tolerance, maximum). The challenge is to select a small subset of the trait variables that explain a large fraction of the variation in the response parameters. The problem is akin to the familiar model selection problem in regression where a response variable is explained by a number of explanatory variables also called covariates (whether continuous or discrete factors). For covariate selection we used the Bayesian variable selection approach of Yuan and Lin (2005). The same approach is also used to find the linear combination of environmental variables that best explains the species data through trait modulated Gaussian logistic response curves.

Bayesian techniques define posterior model probabilities that automatically penalize more complex models, providing a way to select models. Because these probabilities can be very difficult to compute, Bayesian analyses typically use two common approximations, the Bayesian (BIC) and deviance (DIC) information criteria. The BIC is similar to the AIC, and similarly requires an estimate of the number of parameters. The DIC makes weaker assumptions, automatically estimates a penalty for model complexity and is automatically calculated by the WinBUGS/OpenBUGS program (http://www.mrc-bsu.cam.ac.uk/bugs). Despite dispute among statisticians about its properties, the DIC is rapidly gaining popularity among statistician and ecologists.

DIC is used as a preliminary tool for comparing competing models for individual environmental variables and the best linear combination of them (the latent variable). Of course, not all environmental factors are equally important and some factors may perhaps be combined into a synthetic (latent) environmental gradient. The latent variable is formed by a linear combination of environmental variables that are presumed to maximally explain the species distribution.

In Chapter 3, no large difference in the DIC was observed between models with and without traits. Nevertheless, some traits appeared important as judged by their selection by the Yin and Lin approach. The insensitiveness of DIC to the traits is rather unexpected as traits could be usefully selected in chapter 2 on the basis of AIC, the frequentistic equivalent of DIC. The usefulness of DIC for trait selection needs further research.

Chapter 3 fits the Gaussian models in a fully Bayesian framework employing MCMC simulation to generate posterior samples from the joint posterior distribution, which are used to make various posterior inferences. Bayesian methods are easy to implement and provide not only point estimates but also compute confidence intervals for model parameters. The downside is that MCMC simulation may be (computer) time-consuming and that it might be difficult to assess convergence, beyond assessment by Gelman's R-statistic. By contrast, the R software for GLMM that we used (lme4) reports explicitly that its optimization procedure did not converge. Problems with convergence were encountered both using GLMM and the Bayesian approach, but were not so serious as to make the methods impractical.

When a system is described by a statistical model, model complexity leads to a very large computing time and poor estimation, especially if the number of predictors is large relative to the data size. As an alternative to and improvement over stepwise methods, shrinkage methods have been proposed. One of these is the Relevance vector machine (RVM) which has gained popularity within the machine learning community (Tipping 2001, Bishop 2006, Xu 2007) as it does not need cross-validation. RVM introduces a Gaussian prior to regression coefficients, with one variance component for each predictor, which are then estimated by maximizing the marginal likelihood (Type-II maximum likelihood or empirical Bayes). RVM yields variance estimates that are MAP under a uniform prior for the variances.

The Type-II maximum likelihood (Type-II ML) shrinks the coefficients to zero and readily sets some of the coefficients to zero, namely if their variance component is estimated to be zero (Chapter 5). These zero variance coefficients are equivalent to pruning the corresponding predictors from the model. Hence this method readily helps in pruning predictors from the model and does simultaneous variable selection and model estimation.

We studied the selection properties of RVM. RVM selects predictors when the absolute z-ratio (|least squares estimate|/standard error) exceeds 1 in the case of orthogonal predictors and, for two

predictors, this still holds true for correlated predictors when the other z-ratio is large. RVM selects the stronger of two highly correlated predictors (Chapter 5). Compared to other regularization methods e.g. LASSO and PLS, RVM is outcompeted in terms of the prediction performance. The main conclusion from these expressions is again that RVM selects but is also very tolerant in allowing predictors to stay in the model. Despite the fact that RVM has gained popularity in the fields of machine learning, theoretically and empirically, Type-II ML is not the general answer in high dimensional prediction problems.

The empirical Bayes approach in RVM and its extensions by Xu (2010) are not supported as approximations to a fully Bayesian approach (Chapter 5). This fact decreases their credibility; they cannot lend for proven optimality properties of the Bayesian approach.

Community data are multivariate and several environmental factors and traits variables are interacting. Therefore to understand patterns of variation in species density a multivariate approach will be necessary as one-to-one approaches might not be sufficient. This way, selecting important trait-environment interactions becomes a model selection problem in mixed models.

Chapter 6 developed a novel multi-trait and multi-environmental variable model selection method called tiered forward selection. In the first tier, the random factors are selected, in the second, the fixed effects are selected and in the final tier non-significant terms are removed based on a predictive modified Akaike information criterion. Here random factors are the environmental variables while the fixed effects are related to traits and trait-environment interactions. Further we compared the performance of mixed model with the fourth corner method, the linear trait-environment method (LTE) and the one-shot method, namely Type-II maximum likelihood (Type-II ML).

LTE estimates the parameters by least square. In contrast to LTE, linear mixed model (LMM) estimates the fixed effects by generalized least square estimation and their significance can be test by parametric bootstrap (Chapter 6). Type-II ML also identifies the trait-environment interactions which are also identify by other methods but the selection is optimistic. It encountered too many trait-environment interactions and many of them were not common to other methods as Type-II ML is very tolerant in allowing predictors to stay in the model (Chapter 5). The multi-trait multi environment method resulted in less number of significant trait-environment interactions.

The fourth-corner method (Dray and Legendre 2008) test the significance of trait-environment relationship by a permutation test. Dray and Legendre (2008) offer different permutation scenarios but non faithfully controlled the type I error. The linear trait-environment (LTE) method (Cormont et al. 2011) was developed as an alternative to the fourth corner method to account for negative species data values. The method is linear and as such is more closely related to the linear mixed model than the fourth corner method. However, just like the fourth corner method, LTE has no

variance components. The interaction between a trait and an environmental variable in LTE captures the trait-environment relationship. The significance of this interaction is tested by a Monte Carlo permutation test as in the fourth corner problem. The LTE differs from the fourth-corner method by using multivariate linear regression (Cormont et al. 2011).

In this thesis we considered trait-environment relationships. In hindsight, perhaps the most important reason that it is difficult to quantify this relationship is that traits are measured on species and environmental variables on sites. So how can these be related? They can only be related via the sites × species data. So, there is a third entity involved, which complicates the issue.

Statisticians are often keen to distinguish interactions from correlations. Two variables are correlated when a change in one variable is likely to be associated with a change in the other. By contrast, interaction involves a third variable and considers the effects of the variables on this third variable. Two variables are said to interact when the one variable modifies the effect of the other on the third variable. In terms of regression modeling, the third variable is the response variable, the other two are predictor variables and the interaction might be represented by the product of the predictor variables.

Is the trait-environment relationship now best expressed as a correlation as in the fourth corner problem (and LTE) or as an interaction as in the GLMM model? The fourth corner problem is able to express the relationship as a correlation by taking the individual organism as the statistical unit: the cases are the individual organisms. This trick gives the third variable (the sites × species data) another role; the elements become weights. This is *the* logical approach when individuals are (randomly) sampled rather than sites. However, in the practice of much ecological research, primarily sites are sampled and then individuals within sites. The sampling process is thus hierarchical and hierarchical statistical models are thus a natural way to model it. We toke this hierarchical approach in this thesis (Chapter 2). By giving the third variable the role of response variable, the trait-environment relationship becomes naturally an interaction. In contrast with the fourth corner problem, this approach does not ignore the information that some species are absent at a site. The advantage of the GLMM is that it has the potential of predictive use: which species from a species pool are expected to occur under specified environmental conditions when we only know the trait values of the species in the pool. Of course, at the current stage, we are still ignoring any competition and successional processes that must also be important in community assembly, but that does not necessarily invalidate the prediction. It makes it less precise.

Further Research

There has been growing interest in how information about phylogenetic relationships between co-occurring species aids our understanding of community assembly. It should be noted that species are not phylogenetically independent. This seriously precludes claims that the traits we observed as

associated to an environmental variable have evolved independently as an adaptive response to local conditions. In this view, there may also be pseudo-replication on the level of species and not only on the level of the basic observations (presence-absence) as noted in Chapter 2. So far we did neither considered phylogeny, which puts constraints on the way traits may evolve in evolutionary process (Prinzing et al. 2008), nor the spatial configuration of the sites, which set constraints on dispersion (Dray and Legendre 2008, Ozinga et al. 2004). Both aspects can be modeled in a mixed effect models through additional random effects whose correlation depends on either phylogenic association or spatial distance. When these aspects are modeled properly, the GLMM is able to overcome the above-mentioned problems of pseudo-replication. The greatest obstacle for application of this approach is the absence of general software to perform the necessary calculations. In the absence of software, the parameter estimation in these cases is not easy to implement (Bolker et al. 2009). For example, covariance matrices cannot be specified in the GLMM R package lme4. Recently, Ives and Helmus (2011) investigated the phylogenetic structure in community data in combination with either a single environmental variable or a single trait variable or no external data, but not in combination with both a trait and an environmental variable (the case we studied in this thesis).

In an attempt to add the phylogenetic structure in a simpler way, we extended the GLMM analysis by adding principal coordinates of the distance matrix of the phylogenetic tree as nuisance traits. We selected the significant principal coordinates and added those to the GLMM so as to adjust for phylogeny. In the Dune Meadow Data (chapter 2), phylogeny decreased the SigAIC of the model but did not influence the significance of trait-environment interactions. Methods that take an integrative approach to the analysis of traits, phylogeny, environment and spatial configuration merit further research, also in terms of practical software implementation.

This thesis offer new insight in species distribution modeling. The hierarchical approach to estimate trait-environment interactions may have applications in other fields of sciences where one has three table data. One of the promising areas is genetics where the interest is to estimate gene-by-environment (G×E) interactions.

Mixed models are extremely flexible and form a computationally attractive tool for modelling species distribution of complex and large datasets that are common in ecology. They are invaluable when the random variation is the focus of attention, particularly in studies of ecological heterogeneity of species. Fitting models to data is more informative and statistically powerful than informal approaches. In this thesis, we have encouraged ecologists to choose appropriate modelling tools for analyses, and to use them wisely.

References

Abramovich, F., Y. Benjamini, D. L. Donoho, and I. M. Johnstone. 2005. Adapting to unknown sparsity by controlling the false discovery rate. The Annals of Statistics. 34:584-653.

Ackerly, D. D. 2003. Community assembly, niche conservatism, and adaptive evolution in changing environments. International Journal of Plant Sciences 164.

Aho, K. 2011 asbio; a collection of statistical tools for biologists. R package version 0.3-36. http://cran.r-project.org/web/packages/asbio/index.html..

Akaike, H. 1973. Information Theory and an Extension of the Maximum Likelihood Principle. Pages 267-281 in 2nd International Symposium on Information Theory, Tsahkadsor, Armenian SSR

Austin, M. 2007. Species distribution models and ecological theory: A critical assessment and some possible new approaches. Ecological Modelling 200:1-19.

Austin, M. P. 1987. Models for the analysis of species' response to environmental gradients. Plant Ecology 69:35-45.

Bates, D., M. Maechler, and B. Bolker. 2011. lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-39, http://CRAN.R-project.org/package=lme4.

Bayley, P. B. and H. W. Li. 1992 Riverine fishes. Pages 252-281. In P. Calow and G. E. Petts, editors. The Rivers Handbook: Hydrological and Ecological Principles. Vol. 1 Blackwell Scientific Publications, Oxford.

Begon, M. J., L. Harper, and C. R. Townsend. 1990 Ecology: Individuals, populations and communities. Blackwell, Oxford.

Berger, J. 1985. Statistical Decision Theory and Bayesian Analysis. Springer.

Bernhardt-Romermann, M., C. Romermann, R. Nuske, A. Parth, S. Klotz, W. Schmidt, and J. Stadler. 2008. On the identification of the most suitable traits for plant functional trait analyses. Oikos 117:1533-1541.

Bishop, C. 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer, New York.

Bolker, B. M., M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J. S. S. White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology and Evolution 24:127-135.

Bozdogan, H. 1987. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. Psychometrika 52:345-370.

Braun-Blanquet, J. 1964. Pflanzensociologie. 3 edition. Springer, Wein.

Broman, K. W. and T. R. Speed. 2002. A model selection approach for the identification of quantitative trait loci in experimental crosses. Journal of the Royal Statistical Society. Series B: Statistical Methodology 64: 641-656.

Brown, J. H., J. F. Gillooly, A. P. Allen, and V. M. W. Savage, G. B. 2004. Toward a metabolic theory of ecology. Ecology 85:1771- 1789.

Brown, P. J., M. Vannucci, and T. Fearn. 1998. Multivariate Bayesian variable selection and prediction. Journal of the Royal Statistical Society. Series B: Statistical Methodology 60:627-641.

Burnham, K. P. and D. R. Anderson. 2002. Model selection and multimodel inference : a practical information-theoretic approach. Springer, New York.

Burns, C. W. 1968. The relationship between body size of filter-feeding Cladocera and the maximum size of particle ingested. Limnology and Oceanography 13:675-678.

Carroll, C., R. F. Noss, and P. C. Paquet. 2001. Carnivores as focal species for conservation planning in the rocky mountain region. Ecological Applications 11:961-980.

Celeux, G., F. Forbes, C. P. Robert, D. M. Titterington, I. Futurs, and I. Rhône-alpes. 2006. Deviance information criteria for missing data models. Bayesian Analysis 1:651-674.

Chessel, D., A. B. Dufour, and S. Dray. 2011. ade4: Analysis of Ecological Data: Exploratory and Euclidean methods in Environmental sciences. R package version 1.4-17.http://CRAN.R-project.org/package=ade4..

Clements, F. E. 1916. Plant succession: an analysis of the development of vegetation. Carnegie Institute Washington D.C. USA.

Conley, D. J. 2002. Terrestrial ecosystems and the global biogeochemical silica cycle. Global Biogeochemical Cycles 16:1121.

Cormont, A., C. C. Vos, C. A. M. van Turnhout, R. P. B. Foppen, and C. J. F. ter Braak. 2011. Using life-history traits to explain bird population responses to changing weather variability. Climate Research 49:59-71.

Cornwell, W. K. and D. D. Ackerly. 2009. Community assembly and shifts in plant trait distributions across an environmental gradient in coastal California. Ecological Monographs 79:109-126.

Cramer, W. and H. Hytteborn. 1987. The separation of fluctuation and long-term change in vegetation dynamics of a rising seashore. Plant Ecology 69:157-167.

Crawley, J. M. 2002. Statistical Computing: An Introduction to Data Analysis using S-Plus. Jhon Wiley & Sons.

de Rooij, M. 2007. The distance perspective of generalized biadditive models: scalings and transformations Journal of Computational and Graphical Statistics 16: 210-227.

Diamond, J. M. 1975. Assembly of species communities. Harvard University Press, Cambridge, Massachusetts.

Díaz, S., A. Acosta, and M. Cabido. 1992. Morphological Analysis of Herbaceous Communities under Different Grazing Regimes. Journal of Vegetation Science 3:689-696.

Díaz, S., I. Noy-Meir, and M. Cabido. 2001. Can grazing response of herbaceous plants be predicted from simple vegetative traits? Journal of Applied Ecology 38:497-508.

Diggle, P. J., P. Heagerty, K. Y. Liang, and S. L. Zeger. 2002. Analysis of Longitudinal Data. 2nd edition. Oxford University Press, Oxford, U. K.

Dolédec, S. and D. Chessel. 1994. Co-inertia analysis: an alternative method for studying species-environment relationships. Freshwater Biology 31:277-294.

Dolédec, S., B. Statzner, and M. Bournard. 1999. Species traits for future biomonitoring across ecoregions: patterns along a human-impacted river. Freshwater Biology 42:737-758.

Dolédec, S., D. Chessel, C. J. F. ter Braak, and S. Champely. 1996. Matching species traits to environmental variables: a new three-table ordination method. Environmental and Ecological Statistics 3:143-166.

Donoho, D. and J. Johnstone. 1994. Ideal spatial adaptation by wavelet shrinkage. Biometrika 81:425-455.

Donoho, D. L. 1995. De-noising via soft-thresholding. IEEE Transactions on Information Theory 41:613-627

Dray, S. and P. Legendre. 2008. Testing the species traits environment relationships: The fourth-corner problem revisited. Ecology 89:3400-3412.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani. 2004. Least Angle Regression. The Annals of Statistics 32:407-451.

Elliott, J. A., C. S. Reynolds, and A. E. Irish. 2001. An investigation of dominance in phytoplankton using the PROTECH model. Freshwater Biology 46:99-108.

Falkowski, P. G., E. A. Laws, R. T. Barber, and J. W. Murray. 2003. Phytoplankton and their role in primary, new, and export production. Pages 99-121 in M. J. R. Fasham, editor. Ocean biogeochemistry: the role of the ocean carbon cycle in global change. Springer.

Fan, J. and R. Li. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96:1348-1360.

Faul, A. C. and M. E. Tipping. 2002. Analysis of sparse Bayesian learning. Pages 383-389 in T. G. Dietterich, S. Becker, and Z. Ghahramani, editors. Advances in Neural Information Processing Systems 14, Vols 1 and 2.

Fenchel, T. O. and B. J. Finlay. 2004. The Ubiquity of Small Species: Patterns of Local and Global Diversity. BioScience 54:777-784.

Ferber, L., S. Levine, A. Lini, and G. Livingston. 2004. Do cyanobacteria dominate in eutrophic lakes because they fix atmospheric nitrogen? Freshwater Biology 49:690-708.

Fernandez, G. C. 2007. Model Selection in PROC MIXED - A User-Friendly SAS® Macro Application.in SAS Global forum Orlando FL.

Figueiredo, M. A. T. 2003. Adaptive sparseness for supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 25:1150-1159.

Follows, M. J., S. Dutkiewicz, S. Grant, and S. W. Chisholm. 2007. Emergent Biogeography of Microbial Communities in a Model Ocean. Science 315:1843-1846.

Frank, I. E. and J. H. Friedman. 1993. A Statistical View of Some Chemometrics Regression Tools. Technometrics 35:109-135.

Friedman, J. H., T. Hastie, and R. Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33:1-21.

Fuhrman, J. A., J. A. Steele, I. Hewson, M. S. Schwalbach, M. V. Brown, J. L. Green, and J. H. Brown. 2008. A latitudinal diversity gradient in planktonic marine bacteria. Proceedings of the National Academy of Sciences USA 105:7774-7778.

Garnier, E., G. Laurent, A. Bellmann, S. Debain, P. Berthelier, B. Ducout, C. Roumet, and M. L. Navas. 2001. Consistency of Species Ranking Based on Functional Leaf Traits. New Phytologist 152:69-83.

Garwood, N. C. 1996. Functional morphology of tropical tree seedlings. In M. D. Swaine [ed.], The ecology of tropical forest tree seedlings,: 59-129. UNESCO, Paris, France.

Gauch, H. 1982. Multivariate analysis in community ecology. Cambridge University Press, Cambridge.

Gelman, A. and J. Hill. 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

George, E. I. and R. E. McCulloch. 1993. Variable Selection Via Gibbs Sampling. Journal of the American Statistical Association 88:881-889.

Gibson, L. A., B. A. Wilson, D. M. Cahill, and J. Hill. 2004. Spatial prediction of rufous bristlebird habitat in a coastal heathland: a GIS-based approach. Journal of Applied Ecology 41:213-223.

Gideon, S. 1978. Estimating the Dimension of a Model. The Annals of Statistics 6:461-464.

Gillies, C. S., M. Hebblewhite, S. E. Nielsen, M. A. Krawchuk, C. L. Aldridge, J. L. Frair, D. J. Saher, C. E. Stevens, and C. L. Jerde. 2006. Application of random effects to the study of resource selection by animals. Journal of Animal Ecology 75:887-898.

Gimenez, O., S. J. Bonner, R. King, R. A. Parker, S. P. Brooks, L. E. Jamieson, V. Grosbois, B. J. T. Morgan, L. Thomas, D. L. Thomson, E. G. Cooch, and M. J. Conroy. 2009. WinBUGS for Population Ecologists: Bayesian Modeling Using Markov Chain Monte Carlo Methods. Modeling Demographic Processes In Marked Populations. Pages 883-915 in G. P. Patil, editor. Springer US.

Golub, G. H. and C. F. van Loan. 1989. Matrix computations The John Hopkins University Press, Baltimore.

Green, P. J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82:711-732.

Green, P. T. and P. A. Juniper. 2004. Seed Mass, Seedling Herbivory and the Reserve Effect in Tropical Rainforest Seedlings. Functional Ecology 18:539-547.

Green, R. H. 1971. A Multivariate Statistical Approach to the Hutchinsonian Niche: Bivalve Molluscs of Central Canada. Ecology 52:544-556.

Greenacre, M. J. 1984. Theory and applications of correspondence analysis. Academic Press, London.

Greven, S. and T. Kneib. 2010. On the behaviour of marginal and conditional AIC in linear mixed models. Biometrika.

Grime, J. P. 1977. Evidence for the Existence of Three Primary Strategies in Plants and Its Relevance to Ecological and Evolutionary Theory. The American Naturalist 111:1169-1194.

Grime, J. P. 1979. Plant strategies and vegetation processes. Chichester; Wiley.

Grosbois, V., M. P. Harris, T. Anker-Nilssen, R. H. McCleery, D. N. Shaw, B. J. T. Morgan, and O. Gimenez. 2009. Modeling survival at multi-population scales using mark-recapture data. Ecology 90:2922-2932.

Grubb, P. J. 1977. Control of forest growth and distribution on wet tropical mountains: with special reference to mineral nutrition. Annual Review of Ecology and Systematics 8:83-107.

Guisan, A. and N. E. Zimmermann. 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135:147-186.

Guisan, A. and W. Thuiller. 2005. Predicting species distribution: Offering more than simple habitat models. Ecology Letters 8:993-1009.

Gurka, M. J. 2006. Selecting the Best Linear Mixed Model Under REML. The American Statistician 60:19-26.

Hamm, C. E., R. Merkel, O. Springer, P. Jurkojc, C. Maler, K. Prechtel, and V. Smetacek. 2003. Architecture and material properties of diatom shells provide effective mechanical protection. Nature 421:841-843.

He, F. 2010. Maximum entropy, logistic regression, and species abundance. Oikos 119:578-582.

Henry, H., H. Stevens, D. E. Bunker, S. A. Schnitzer, and W. P. Carson. 2004 Establishment limitation reduces species recruitment and species richness as soil resources rise. Journal of Ecology 92 339-347.

Hill, M. O. and H. G. Gauch. 1980. Detrended correspondence analysis: An improved ordination technique. Plant Ecology 42:47-58.

Hoerl, A. E. and R. W. Kennard. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 12:55-67.

Hoeting, J. A., R. A. Davis, A. A. Merton, and S. E. Thompson. 2006. Model Selection For Geostatistical Models. Ecological Applications 16:87-98.

Hosmer, W. D. and S. Lemeshow. 2000. Applied Logistic Regression, 2nd Edition. Wiley, New York.

Huang, Y., H. Wu, J. Holden-Wiltse, and E. P. Acosta. 2011 A dynamic Bayesian nonlinear mixed-effects model of HIV response incorporating medication adherence, drug resistance and covariates. The Annals of Applied Statistics 5:551-577.

Hubbell, S. P. 2001. The Unified Neutral Theory of Biodiversity and Biogeography. Princeton University Press, Princeton.

Hurlbert, S. H. 1984. Pseudoreplication and the Design of Ecological Field Experiments. Ecological Monographs 54:187-211.

Hurvich, C. M. and C. Tsai. 1989. Regression and time series model selection in small samples. Oxford University Press.

Hutchinson, G. E. 1961. The paradox of plankton. The American Naturalist 882:137-145.

Ihm, P. and H. Van Groenewoud. 1984. Correspondence analysis and Gaussian ordination. COMPSTAT Lectures:5-60.

Ives, A. R. and M. R. Helmus. 2011. Generalized linear mixed models for phylogenetic analyses of community structure. Ecological Monographs 81:511-525.

Jamil, T., K. Carla, and C. J. F. ter Braak. in prep. A unimodal species response model relating traits to environment with application to phytoplankton communities

Jari, O., F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, and H. Wagner. 2011 vegan: Community Ecology R Package version 1.17-12. http://cran.r-project.org/web/packages/vegan/index.html.

Johnson, C. J., D. R. Seip, and M. S. Boyce. 2004. A quantitative approach to conservation planning: Using resource selection functions to map the distribution of mountain caribou at multiple spatial scales. Journal of Applied Ecology 41:238-251.

Johnstone, I. M. and B. W. Silverman. 2004. Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. Annals of Statistics 32:1594-1649.

Jongman , R. H. G., C. J. F. ter Braak, and O. F. R. van Tongeren. 1995. Data Analysis in Community and Landscape Ecology. Cambridge University Press.

Kahmen, S. and P. Poschlod. 2004. Plant functional trait responses to grassland succession over 25 years. Journal of Vegetation Science 15:21-32.

Kahmen, S. and P. Poschlod. 2008. Effects of grassland management on plant functional trait composition. Agriculture Ecosystems and Environment 128:137-145.

Karatzoglou, K., A. Smola, K. Hornik, and A. Zeileis. 2004 kernlab - An S4 Package for Kernel Methods in R. Journal of Statistical Software 11 1-9.

Keddy, P. A. 1992. A pragmatic approach to functional ecology. Functional Ecology 6:621-626.

Keddy, P. A., L. Twolan-Strutt, and I. C. Wisheu. 1994. Competitive Effect and Response Rankings in 20 Wetland Plants: Are They Consistent Across Three Environments? Journal of Ecology 82:635-643.

Kirk, J. T. O. 1996. Light and photosynthesis in aquatic ecosystems 2nd Edition. Cambridge University Press, Cambridge.

Kleyer, M., R. M. Bekker, I. C. Knevel, J. P. Bakker, K. Thompson, M. Sonnenschein, P. Poschlod, J. M. van Groenendael, L. Klimes, J. Klimesova, S. Klotz, G. M. Rusch, M. Hermy, D. Adriaens, G. Boedeltje, B. Bossuyt, A. Dannemann, P. Endels, L. Gotzenberger, J. G. Hodgson, A. K. Jackel, I. Kuhn, D. Kunzmann, W. A. Ozinga, C. Romermann, M. Stadler, J. Schlegelmilch, H. J. Steendam, O. Tackenberg, B. Wilmann, J. H. C. Cornelissen, O. Eriksson, E. Garnier, and B. Peco. 2008. The LEDA Traitbase: a database of life-history traits of the Northwest European flora. Journal of Ecology 96:1266-1274.

Klimešová, J. and L. Klimeš. 2006. CLO-PLA3. a database of clonal growth architecture of Central European plants. http://clopla.butbn.cas.cz/.

Kosten, S., G. Lacerot, E. Jeppesen, D. da Motta Marques, E. H. van Nes, N. Mazzeo, and M. Scheffer. 2009b. Effects of submerged vegetation on water clarity across climates. Ecosystems 12:1117-1129.

Kosten, S., V. L. M. H. Huszar, E. Bécares, L. S. Costa, E. van Donk, L.-A. Hansson, E. Jeppesen, C. Kruk, G. Lacerot, N. Mazzeo, L. De Meester, B. Moss, M. Lürling, T. Nõges, S. Romo, and M. Scheffer. 2011. Warmer climates boost cyanobacterial dominance in shallow lakes. Global Change Biology. doi: 10.1111/j.1365-2486.2011.02488.x.

Kosten, S., V. L. M. Huszar, N. Mazzeo, M. Scheffer, L. S. L. Sternberg, and E. Jeppesen. 2009a. Limitation of phytoplankton growth in South America: no evidence for increasing nitrogen limitation towards the tropics. Ecological Applications 19:1791-1804.

Kotowski, W., O. Beauchard, W. Opdekamp, P. Meire, and R. Van Diggelen. 2010. Waterlogging and canopy interact to control species recruitment in floodplains. Functional Ecology 24:918-926.

Kruk, C. 2010. Morphology Captures Function in Phytoplankton. A Large-Scale Analysis of Phytoplankton Communities in Relation to their Environment. Wageningen University Wageningen.

Kruk, C., E. T. H. M. Peeters, E. H. Van Nes, V. L. M. Huszar, L. S. Costa, and M. Scheffer. 2011. Phytoplankton Community Composition can be Predicted Best in Terms of Morphological Groups. Limnology and Oceanography 56:110-118.

Kruk, C., N. Mazzeo, G. Lacerot, and C. S. Reynolds. 2002. Classification schemes for phytoplankton: a local validation of a functional approach to the analysis of species temporal replacement. Journal of Plankton Research 24:901-912.

Kruk, C., V. L. M. Huszar, E. T. H. M. Peeters, S. Bonilla, L. Costa, M. Lürling, C. S. Reynolds, and M. Scheffer. 2010. A morphological classification capturing functional variation in phytoplankton. Freshwater Biology 55:614-627.

Kühn, I., W. Durka, and S. Klotz. 2004. BiolFlor a new plant-trait database as a tool for plant invasion ecology. Diversity and Distributions 10:363-365.

Kuhn, M. 2008. Building Predictive Models in R Using the caret Package. Journal of Statistical Software 28:1-26.

Kuhn, M. 2011. caret: Classification and Regression Training. R package version 4.98. http://CRAN.R-project.org/package=caret.

Kuo, L. and B. Mallick. 1998. Variable Selection for Regression Models. Sankhy?: The Indian Journal of Statistics, Series B 60:65-81.

Lahoz-Monfort, J. J., B. J. T. Morgan, M. P. Harris, S. Wanless, and S. N. Freeman. 2011. A capture-recapture model for exploring multi-species synchrony in survival. Methods in Ecology and Evolution 2:116-124.

Lampert, W. 1987. Feeding and nutrition in Daphnia. Memorie dell' Istituto Italiano di Idrobiologia 45:143-192.

Lampert, W. and U. Sommer. 2007. Limnoecology. 2nd edition. Oxford University Press, Oxford.

Lavorel, S. and E. Garnier. 2002. Predicting changes in community composition and ecosystem functioning from plant traits: revisiting the Holy Grail. Functional Ecology 16:545-556.

Legendre, P., R. G. Galzin, and M. L. Harmelin-Vivien. 1997. Relating behavior to habitat: Solutions to the fourth-corner problem. Ecology 78:547-562.

Lehman, J. T. 1988. Selective herbivory and its role in the evolution of phytoplankton growth strategies. Pages 369-387 Growth and reproductive strategies of freshwater phytoplankton.

Lenssen, J. P. M., F. B. J. Menting, and W. H. V. d. Putten. 2003. Plant Responses to Simultaneous Stress of Waterlogging and Shade: Amplified or Hierarchical Effects? New Phytologist 157:281-290.

Lewis, W. M. J. 1976. Surface/volume ratio: implications for phytoplankton morphology. Science 192:885-887.

Li, Y., C. Campbell, and M. Tipping. 2002. Bayesian automatic relevance determination algorithms for classifying gene expression data. Bioinformatics 18:1332-1339.

Litchman, E. and C. A. Klausmeier. 2008. Trait-Based Community Ecology of Phytoplankton. Annual Review of Ecology Evolution and Systematics 39:615-639.

Littell, R. C., G. A. Milliken, W. W. Stroup, R. D. Wolfinger, and O. Schabenberger. 2006 SAS for Mixed Models. 2nd edition. SAS Institute Inc, Cary NC

Liu, K., R. J. Eastwood, S. Flynn, Turner, R.M., and W. H. Stuppy. 2008 Seed Information Database. Royal Botanic Gardens, Kew.

MacKay, D. J. C. 1992. The Evidence Framework Applied to Classification Networks. Neural Computation 4:720-736.

Malthus, T. J. and S. F. Mitchell. 2006. On the ocurrence, causes and potential consequences of low zooplankton to phytoplankton ratios in New Zealand lakes. Freshwater Biology 22:383-394.

Margalef, R. 1978. Life-forms of phytoplankton as survival alternatives in an unstable environment. Oceanologica Acta 1:493-509.

McGill, B. J., B. J. Enquist, E. Weiher, and M. Westoby. 2006. Rebuilding community ecology from functional traits. Trends in Ecology and Evolution 21:178-185.

Menezes, S., D. J. Baird, and A. Soares. 2010. Beyond taxonomy: a review of macroinvertebrate trait-based community descriptors as tools for freshwater biomonitoring. Journal of Applied Ecology 47:711-719.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.

Miller, A. J. 2002 Subset Selection in Regression Chapman & Hall, London.

Minchin, P. R. 1989. Montane vegetation of the Mt. Field massif, Tasmania: a test of some hypotheses about properties of community patterns. Plant Ecology 83:97-110.

Moss, B., S. Kosten, M. Meerhoff, R. Battarbee, E. Jeppesen, N. Mazzeo, K. Havens, G. Lacerot, Z. Liu, L. de Meester, H. Paerl, and M. Scheffer. 2011. Allied attack: climate change and eutrophication. Inland Waters 1:101-105.

Naselli-Flores, L. and R. Barone. 2007. Pluriannual morphological variability of phytoplankton in a highly productive Mediterranean reservoir (Lake Arancio, Southwestern Sicily). Hydrobiologia 578:87-95.

Neal, R. 1996. Bayesian Learning for Neural Networks (Lecture Notes in Statistics). Springer.

Nelder, J. A. 2008. What is the Mixed-Models Controversy? International Statistical Review 76:134-135.

O'Hagan, A. and J. Forster. 2004. Kendall's advanced theory of statistics: Bayesian inference. Arnold, London.

O'Hara, R. B. and M. J. Sillanpää. 2009 A review of Bayesian variable selection methods: what, how and which. Bayesian Analysis 4 85-118.

Økland, R. H. 1986. Reseating of ecological gradients. II. The effect of scale on symmetry of species response curves. Nordic Journal of Botany 6:661-670.

Oksanen, J. and P. R. Minchin. 2002. Continuum theory revisited: what shape are species responses along ecological gradients? Ecological Modelling 157:119-129.

Oksanen, J., E. Läärä, K. Tolonen, and B. G. Warner. 2001. Confidence Intervals for the Optimum in the Gaussian Response Function. Ecology 82:1191-1197.

Oksanen, J., E. Läärä, P. Huttunen, and J. Meriläinen. 1988. Estimation of pH optima and tolerances of diatoms in lake sediments by the methods of weighted averaging, least squares and maximum likelihood, and their use for the prediction of lake acidity. Journal of Paleolimnology 1:39-49.

Ozinga, W. A., J. H. J. Schaminée, R. M. Bekker, S. Bonn, P. Poschlod, O. Tackenberg, J. Bakker, and J. M. v. Groenendael. 2005b. Predictability of plant species composition from environmental conditions is constrained by dispersal limitation. Oikos 108:555-561.

Ozinga, W. A., R. M. Bekker, J. H. J. Schaminée, and J. M. Van Groenendael. 2004. Dispersal potential in plant communities depends on environmental conditions. Journal of Ecology 92:767-777.

Ozinga, W. A., S. M. Hennekens, J. H. J. Schaminée, R. M. Bekker, A. Prinzing, S. Bonn, P. Poschlod, O. Tackenberg, K. Thompson, J. P. Bakker, and J. M. v. Groenendael. 2005a. Assessing the Relative Importance of Dispersal in Plant Communities Using an Ecoinformatics Approach. Folia Geobotanica 40:53-67.

Padisák, J., E. Soróczki-Pinté and Z. Rezner. 2003. Sinking properties of some phytoplankton shapes and the relation of form resistance to morphological diversity of phytoplankton--an experimental study. Hydrobiologia 500:243-257.

Paerl, H. and J. Huisman. 2008. Blooms Like It Hot. Science 320:57-58.

Paerl, H. W. and J. Huisman. 2009. Minireview Climate change: a catalyst for global expansion of harmful cyanobacterial blooms. Environmental Microbiology Reports. 1 1:27-37.

Palmer, M. W. and P. M. Dixon. 1990. Small-Scale Environmental Heterogeneity and the Analysis of Species Distributions along Gradients. Journal of Vegetation Science 1:57-65.

Park, T. and G. Casella. 2008. The Bayesian Lasso. Journal of the American Statistical Association 103:681-686.

Paterson, S. and J. Lello. 2003. Mixed models: Getting the best use of parasitological data. Trends in Parasitology 19:370-375.

Pearsall, W. H. 1950 Mountains and Moorlands, London.

Peretyatko, A., S. Teissier, J.-J. Symoens, and L. Triest. 2007. Phytoplankton biomass and environmental factors over a gradient of clear to turbid peri-urban ponds. Aquatic Conservation: Marine and Freshwater Ecosystems 17:584-601.

Pinheiro, J. C. and D. M. Bates. 2000. Mixed-Effects Models in S and SPLUS. Springer.

Poff, N. L., J. D. Olden, N. K. M. Vieira, D. S. Finn, M. P. Simmons, and B. C. Kondratieff. 2006. Functional Trait Niches of North American Lotic Insects: Traits-based ecological

applications in light of phylogenetic relationships. Journal of the North American Benthological Society 25:730-755.

Polson, N. G. and J. G. Scott. 2009. Alternative global-local shrinkage priors using hypergeometric-beta mixtures. Technical report. University of Chicago, DOI 10.1.1.161.3592, Chicago.

Pöyry, J., M. Luoto, R. K. Heikkinen, and K. Saarinen. 2008. Species traits are associated with the quality of bioclimatic models. Global Ecology and Biogeography 17:403-414.

Prinzing, A., R. Reiffers, W. G. Braakhekke, S. M. Hennekens, O. Tackenberg, W. A. Ozinga, J. H. J. Schaminée, and J. M. Van Groenendael. 2008. Less lineages - more trait variation: phylogenetically clustered plant communities are functionally more diverse. Ecology Letters 11:809-819.

R Development Core Team. 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing. www.R-project.org, Vienna.

Raxworthy, C. J., E. Martinez-Meyer, N. Horning, R. A. Nussbaum, G. E. Schneider, M. A. Ortega-Huerta, and A. T. Peterson. 2003. Predicting distributions of known and unknown reptile species in Madagascar. Nature 426:837-841.

Reynolds, C. 1987. The response of phytoplankton communities to changing lake environments. Aquatic Sciences - Research Across Boundaries 49:220-236.

Reynolds, C. S. 1984a. The Ecology of Freshwater Phytoplankton. Cambridge University Press, Cambridge.

Reynolds, C. S. 1984b. Phytoplankton periodicity: the interaction of form, function and environmental variability. Freshwater Biology 14:111-142.

Reynolds, C. S. 1988. Functional morphology and the adaptive strategies of freshwater phytoplankton. Pages 388-433 in C. D. Sandgren, editor. Growth and reproductive strategies of freshwater phytoplankton. Cambridge University Press, New York.

Reynolds, C. S. 1997. Vegetation Process in the pelagic: a model for ecosystem theory. Excellence in Ecology. Ecology Institute.

Reynolds, C. S. 1998. What factors influence the species composition of phytoplankton in lakes of different trophic status? Hydrobiologia 369/370:11-26.

Reynolds, C. S. 2006. Ecology of phytoplankton. Cambridge University Press, Cambridge.

Reynolds, C. S. 2007. Variability in the provision and function of mucilage in phytoplankton: facultative responses to the environment. Hydrobiologia 578:37-45.

Reynolds, C. S. and A. E. Irish. 1997. Modelling phytoplankton dynamics in lakes and reservoirs: the problem of in-situ growth rates. Hydrobiologia 349:5-17.

Reynolds, C. S., Jawroski, G. H. M., Cmieche, H.A. and Leedale, G.F. 1981. On the annual cycle of the blue-green alga M. aeruginosa Kütz. Emend. Elenkin. Phil. Trans. R. Soc. London. B 293:419-477.

Reynolds, C., V. Huszar, C. Kruk, L. Naselli-Flores, and S. Melo. 2002. Towards a functional classification of the freshwater phytoplankton. Journal of Plankton Research 24:417-428.

Rice, J., R. D. Ohmart, and B. W. Anderson. 1983. Habitat Selection Attributes of an Avian Community: A Discriminant Analysis Investigation. Ecological Monographs 53:263-290.

Rogers, S. and M. Girolami. 2005. A Bayesian regression approach to the inference of regulatory networks from gene expression data. Bioinformatics 21:3131-3137.

Roweis, S. 1999. Matrix identities. http://www.cs.nyu.edu/~roweis/notes/matrixid.pdf.

Rue, H. and L. Held. 2005 Gaussian Markov Random Fields: Theory and Applications. London: Chapman and Hall-CRC Press.

Ruhland, K., A. M. Paterson, and J. P. Smol. 2008. Hemispheric-scale patterns of climate-related shifts in planktonic diatoms from North American and European lakes. Global Change Biology 14:2740-2754.

Scheffer, M. and E. H. van Nes. 2006. Self-organized similarity, the evolutionary emergence of groups of similar species. Proceedings of the National Academy of Sciences USA 103:6230-6235.

Searle, S. R., G. Casella, and C. E. McCulloch. 2008. Variance components. John Wiley, New York.

Segura, A. M., D. Calliari, C. Kruk, D. Conde, S. Bonilla, and H. Fort. 2011. Emergent neutrality drives phytoplankton species coexistence. Proceedings of the Royal Society B 278:2355-2361

Shipley, B., D. Vile, and Ã. Garnier. 2006. From plant traits to plant communities: A statistical mechanistic approach to biodiversity. Science 314:812-814.

Shipley, B., P. A. Keddy, D. R. J. Moore, and K. Lemky. 1989. Regenration and establishment strategies of emergent macrophytes. Journal of Ecology 77:1093-1110.

Smetacek, V. 2001. A watery arms race. Nature 411:745-745.

Smith, V. H. 2006. Responses of estuarine and coastal marine phytoplankton to nitrogen and phosphorus enrichment. Limnol. Oceanogr. 51:377-384.

Smith, V. H. 2007. Microbial diversity-productivity relationships in aquatic ecosystems. FEMS Microbiology Ecology 62:181-186.

Smol, J. P., A. P. Wolfe, H. J. B. Birks, M. S. V. Douglas, V. J. Jones, A. Korhola, R. Pienitz, K. Rühland, S. Sorvari, D. Antoniades, S. J. Brooks, M.-A. Fallu, M. Hughes, B. E. Keatley, T. E. Laing, N. Michelutti, L. Nazarova, M. Nyman, A. M. Paterson, B. Perren, R. Quinlan, M. Rautio, É. Saulnier-Talbot, S. Siitonen, N. Solovieva, and J. Weckström. 2005. Climate-driven regime shifts in the biological communities of arctic lakes. Proceedings of the National Academy of Sciences of the USA 102:4397-4402.

Sommer, U. 1989. Plankton Ecology: Succession in plankton communities. Springer-Verlag, Berlin.

Sonnier, G., B. Shipley, and M.-L. Navas. 2010. Quantifying relationships between traits and explicitly measured gradients of stress and disturbance in early successional plant communities. Journal of Vegetation Science 21:1014-1024.

Southwood, T. R. E. 1977. Habitat, the Templet for Ecological Strategies? Journal of Animal Ecology 46:337-365.

Southwood, T. R. E. 1988. Tactics, Strategies and Templets. Oikos 52:3-18.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde. 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society. Series B: Statistical Methodology 64:583-639.

Statzner, B., S. Dolédec, and B. Hugueny. 2004. Biological trait composition of European stream invertebrate communities: assessing the effects of various trait filter types. Ecography 27:470-488.

Statzner, B., V. H. Resh, and S. Dolédec. 1994. Ecology of the Upper Rhône River: a test of habitat templet theories. . Special issue of Freshwater Biology 31:253-554.

Stockey, A. and R. Hunt. 1994. Predicting Secondary Succession in Wetland Mesocosms on the Basis of Autecological Information on Seeds and Seedlings. Journal of Applied Ecology 31:543-559.

Stroup, W. 2011. Living with Generalized Linear Mixed Models.in Proceedings of the SAS Global Forum 2011 Conference, Las Vegas Nevada.

Sturtz, S., U. Ligges, and A. Gelman. R2OpenBUGS: A Package for Running OpenBUGS from R

Tansley, A. G. 1939 The British Islands and their Vegetation. Cambridge University Press, Cambridge.

ter Braak, C. J. F. 1987. Unimodal models to relate species to environment. DLO-Agricultural Mathematics Group, 1996: Wageningen, the Netherlands

ter Braak, C. J. F. 1988. Partial canonical correspondence analysis. Pages 551-558 Classification and related methods of data analysis. (ed H. H. Bock), North-Holland, Amsterdam.

ter Braak, C. J. F. 2006. Bayesian sigmoid shrinkage with improper variance priors and an application to wavelet denoising. Comput. Stat. Data Anal. 51:1232-1242.

ter Braak, C. J. F. 2009. Regression by L1 regularization of smart contrasts and sums (ROSCAS) beats PLS and elastic net in latent variable model. Journal of Chemometrics 23:217-228.

ter Braak, C. J. F. and C. W. N. Looman. 1986. Weighted averaging, logistic regression and the Gaussian response model. Plant Ecology 65:3-11.

ter Braak, C. J. F. and I. C. Prentice. 2004. A theory of gradient analysis. Advances in Ecological Research 34:235-282.

ter Braak, C. J. F. and P. F. M. Verdonschot. 1995. Canonical correspondence analysis and related multivariate methods in aquatic ecology. Aquatic Sciences 57:255-289.

ter Braak, C. J. F. and P. Smilauer. 1998. CANOCO Reference Manual and User's Guide to Canoco for Windows: Software for Canonical Community Ordination (version 4). Microcomputer Power, Ithaca, NY, USA.

ter Braak, C. J. F., M. P. Boer, and M. C. A. M. Bink. 2005. Extending Xu's Bayesian Model for Estimating Polygenic Effects Using Markers of the Entire Genome. Genetics 170:1435-1438.

ter Braak, C. J. F., S. Juggins, H. J. B. Birks, and H. van der Voet. 1993. Weighted averaging partial least squares regression (WA-PLS): definition and comparison with other methods for species-environment calibration. Pages 525-560 Chapter 25 in: Multivariate Environmental Statistics, G. P. Patil, and C. R. Rao (eds). Amsterdam: Elsevier (North-Holland).

Thuiller, W. 2007. Biodiversity: Climate change and the ecologist. Nature 448:550-552.

Thuiller, W., D. M. Richardson, P. Pyssek, G. F. Midgley, G. O. Hughes, and M. Rouget. 2005. Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. Global Change Biology 11:2234-2250.

Tibshirani, R. 1996. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B: Statistical Methodology 58:267-288.

Tilman, D., Kiesling, R., Sterner, R. and S. Tilman. 1986. Green, blue-green and diatom algae: taxonomic differences in competetive ability for phosphorus, silicon, and nitrogen. Arch. Hydrobiol. 106:473-485.

Tipping, M. E. 2001. Sparse bayesian learning and the relevance vector machine. J. Mach. Learn. Res. 1:211-244.

Tipping, M. E. and A. Faul. 2003. Fast marginal likelihood maximisation for sparse Bayesian models.in C. M. Bishop and B. J. Frey, editors. Proceedings of the Ninth International Workshop on Arficial Intelligence and Statistics, Key West, FL.

Townsend, C. R. and A. G. Hildrew. 1994. Species traits in relation to a habitat templet for river systems. Freshwater Biology 31:265-275.

Townsend, C., S. DolÉDec, and M. Scarsbrook. 1997. Species traits in relation to temporal and spatial heterogeneity in streams: a test of habitat templet theory. Freshwater Biology 37:367-387.

Van Buuren, S. and K. Groothuis-Oudshoorn. 2011. MICE: Multivariate Imputation by Chained Equations. R package version 2.9. http://www.stefvanbuuren.nl; http://www.multiple-imputation.com.

van Duuren, L., J. Eggink G, J. Kalkhoven, J. Notenboom, A. J. van Strien, and R. Wortelboer. 2003 Biobase CBS, Voorburg/Heerlen, NL.

Verbeke, G. and G. Molenberghs. 2000. Linear Mixed Models for Longitudinal Data. Springer, New York.

Vergnon, R., N. K. Dulvy, and R. P. Freckleton. 2009. Niches versus neutrality: uncovering the drivers of diversity in a species-rich community. Ecology Letters 12:1079-1090.

Vile, D., B. Shipley, and E. Garnier. 2006. Ecosystem productivity can be predicted from potential relative growth rate and species abundance. Ecology Letters 9:1061-1067.

Violle, C. and L. Jiang. 2009. Towards a trait-based quantification of species niche. Journal of Plant Ecology 2:87-93.

Violle, C., M. L. Navas, D. Vile, E. Kazakou, I. H. Fortunel, and E. Garnier. 2007. Let the concept of trait be functional! Oikos 116:882-892.

Wagner, H. H. and M. J. Fortin. 2005. Spatial analysis of landscapes: Concepts and statistics. Ecology 86:1975-1987.

Walker, S. and D. Jackson. 2011. Random-effects ordination: describing and predicting multivariate correlations and co-occurrences. Ecological Monographs.

Wehrens, R. and B.-H. Mevik. 2006. The PLS package 2.0-0. Multivariate regression by partial least squares regression (PLSR) and principal component regression (PCR). http://cran.r-project.org/doc/packages/.

Weiher, E. and P. A. Keddy.1995.The Assembly of Experimental Wetland Plant Communities. Oikos 73:323-335.

Weiher, E., A. van der Werf, K. Thompson, M. Roderick, E. Garnier, and O. Eriksson. 1999. Challenging Theophrastus: A common core list of plant traits for functional ecology. Journal of Vegetation Science 10:609-620.

Weiher, E., G. D. P. Clarke, and P. A. Keddy. 1998. Community Assembly Rules, Morphological Dispersion, and the Coexistence of Plant Species. Oikos 81:309-322.

Weithoff, G. 2003. The concepts of 'plant functional types' and 'functional diversity' in lake phytoplankton - a new understanding of phytoplankton ecology? Freshwater Biology 48:1669-1675.

West, B. T., K. B. Welch, and A. T. Galecki. 2006. Linear Mixed Models: A Practical Guide Using Statistical Software. Chapman & Hall/CRC.

Westoby, M., E. Jurado, and M. Leishman. 1992. Comparative evolutionary ecology of seed size. Trends in Ecology and Evolution 7:368-372.

Whittaker, R. H. 1960. Vegetation of the Siskiyou Mountains, Oregon and California. Ecological Monographs 30:279-338.

Wiens, J. A. 1991. Ecomorphological Comparisons of the Shrub-Desert Avifaunas of Australia and North America. Oikos 60:55-63.

Winder, M., J. E. Reuter, and S. G. Schladow. 2009. Lake warming favours small-sized planktonic diatom species. Proceedings of the Royal Society B 276:427-435.

Wold, S., M. Sjöström, and L. Eriksson. 2001. PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 58:109-130.

Wolfinger, R. 1993. Covariance structure selection in general mixed models. Communications in Statistics - Simulation and Computation 22:1079 - 1106.

Wolfinger, R. D. 1996. Heterogeneous Variance: Covariance Structures for Repeated Measures. Journal of Agricultural, Biological, and Environmental Statistics 1:205-230.

Xu, S. 2007. An Empirical Bayes Method for Estimating Epistatic Effects of Quantitative Trait Loci. Biometrics 63:513-521.

Yuan, M. and Y. Lin. 2005. Efficient Empirical Bayes Variable Selection and Estimation in Linear Models. Journal of the American Statistical Association 100:1215-1225.

Zou, H. and T. Hastie. 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society. Series B: Statistical Methodology 67:301-320.

Zuur, A. F., E. N. Ieno, N. Walker, A. A. Saveliev, and G. M. Smith. 2009. Mixed Effects Models and Extensions in Ecology with R. Springer, Berlin.

Summary

In the last two decades, the interest of community ecologists in trait-based approaches has grown dramatically and these approaches have been increasingly applied to explain and predict response of species to environmental conditions. A variety of modelling techniques are available. The dominant technique is to cluster the species based on their functional traits and then summarize the response of the clusters to environmental change. In general, fitting explicit models to data is always more informative and powerful than more informal approaches.

The central theme of the thesis is how to quantify the relation of traits with the environment using three data tables, data on species occurrence and abundance in sites, data on traits of species and data on the environmental characteristics of sites. In this thesis, we place the challenge of quantifying trait-environment relationships in the context of species distribution modelling, so in the context of species-environment relationships. We present a hierarchal statistical approach to species distribution modelling that efficiently utilize the trait information and that is able to automatically select the relevant traits and environmental characteristics. This model-based approach, coupled with recent statistical developments and increased computing power, opens up possibilities that were unimaginable before.

In **Chapter 2** a hierarchical statistical approach is introduced for modeling and explaining species response along environmental gradients by species traits. The model is an extension of the generalized linear model with random terms that express the between-species variation in response to the environment. This so-called generalized linear mixed model (GLMM) is derived by integrating a two-step procedure into one. As the basic GLMM we take the random intercept and random slope model. To introduce traits, the regression parameters (intercept and slope) are made linearly dependent on the species traits. As a consequence the trait-environment relationship is represented as an interaction term in the model. The method is illustrated using the famous Dune Meadow Data using Ellenberg indicator values as species traits.

Niche theory proclaims that species response to environmental gradients is nonlinear. Each species has preferred an environmental condition in which it can survive and reproduce optimally. Thus each species tends to be most abundant around a specific environmental optimum and the distribution of species along any environmental gradient is usually unimodal, with the maximum at some ecological optimum. For presence-absence data, the simplest unimodal (non-negative) species response curve is the Gaussian logistic response curve with three parameters that characterize the niche: optimum (niche centre), tolerance (niche width) and maximum (expected occurrence at the centre). Niches of species differ between species and species are assumed to be evolutionary adapted. It is difficult to fit the Gaussian logistic model with linear trait submodels

for the parameters with the available (generalized) nonlinear mixed model software. In **Chapter 3** we develop the trait-modulated Gaussian logistic model in which the niche parameters are made linearly dependent on species traits. The model is fitted to data in the Bayesian framework using OpenBUGS (Bayesian inference Using Gibbs Sampling). A Bayesian variable selection method is used to identify which species traits and environmental variables best explain the species data through this model. We extended the approach to find the best linear combination of environmental variables.

In **Chapter 4** we explained why and when (generalized) linear mixed models can effectively analyse unimodal data and presented a graphical tool and statistical test to test for unimodality while fitting just a generalized linear mixed model without any squared or other polynomial term. A GLMM is, of course, a linear model. Despite this fact, it can be used to detect unimodality and to fit unimodal data, with the provision that the differences in niche widths among species are not too large. As graphical tool we suggested to plot the random site effects against the environmental variable. There is an indication for unimodality, when this graph shows a quadratic relationship. The efficacy of GLMM to analyse unimodal data is illustrated by comparing the GLMM approach with an explicit unimodal model approach on simulated data and real data that show unimodality.

When a system is described by a statistical model, model complexity leads to a very large computing time and poor estimation, especially if the number of predictors is large relative to the data size. As an alternative to and improvement over stepwise methods, shrinkage methods have been proposed. One of these is the Relevance vector machine (RVM). RVM assigns individual precisions to weights of predictors which are then estimated by maximizing the marginal likelihood (Type-II ML or empirical Bayes). In **Chapter 5** we investigated the selection properties of RVM both analytically and by experiments. We found that RVM is rather tolerant for predictors to stay in the model and concluded that RVM is not a real solution in high-dimensional data problems.

Chapter 6 further developed the multi-trait and multi-environmental variable model selection method that used Chapter 2 in a linear mixed model context. The method is called tiered forward selection. In the first tier, the random factors are selected, in the second, the fixed effects are selected and in the final tier non-significant terms are removed based on a modified Akaike information criterion. The linear mixed model with the tiered forward selection is compared with Type-II ML and existing methods for detecting trait-environment relationships that are not based on mixed models, namely the fourth corner method and the linear trait-environment method (LTE).

Chapter 7 summarizes the findings of the methods. The limitations of methods are discussed and future research directions in species distribution modeling and integration of these methods in other disciplines are proposed.

Samenvatting

De belangstelling van ecologen voor de rol van kenmerken van soorten in hun verspreiding is de afgelopen twintig jaar enorm toegenomen. Kenmerken worden meer en meer gebruikt om de reactie van soorten op milieufactoren te verklaren en te voorspellen. Een aantal modelleertechnieken zijn hiervoor beschikbaar. De meest gebruikte techniek is om soorten te clusteren op basis van hun functionele kenmerken en om dan de reactie van de clusters op milieuverandering samen te vatten. Over het algemeen is het aanpassen van expliciete modellen aan gegevens altijd informatiever en krachtiger dan een informelere aanpak.

Het centrale thema van het proefschrift is de vraag hoe the relatie tussen kenmerken en milieu te kwantificeren op basis van drie tabellen van gegevens, gegevens over het voorkomen en/of de abundantie van soorten op monsterplekken, gegevens over kenmerken van soorten en gegevens over milieukarakteristieken op de monsterplekken. In dit proefschrift plaatsen we de uitdaging om kenmerk-milieu relaties te kwantificeren in de context van het modelleren van de verspreiding van soorten, dus in de context van soort-milieu relaties. Deze op modellen gebaseerd aanpak, samen met recente ontwikkelingen in de statistiek en de toegenomen kracht van computers, biedt nieuwe mogelijkheden die tevoren ondenkbaar waren.

In hoofdstuk 2 wordt een hiërarchische statistische aanpak voorgesteld om de reactie van soorten op milieugradienten te modeleren en te verklaren op basis van kenmerken van soorten. Het model is een uitbreiding van het gegeneraliseerde lineaire model met random termen die de tussen-soortsvariatie in reaktie op het milieu uitdrukken. Dit generaliseerde lineaire *gemengde* model (GLMM) wordt afgeleid door de twee stappen van een twee-staps procedure te integreren. Als basis GLMM nemen we het random intercept en random helling model. Kenmerken worden aan het model toegevoegd door de regressie parameters (intercept en helling) lineair afhankelijk te maken van de kenmerken van de soorten. Als gevolg hiervan wordt de kenmerk-milieu relatie weergegeven door een interactieterm in het model. Het model wordt geïllustreerd aan de hand van de beroemde Duinweidengegevens met gebruikmaking van de Ellenberg indicatiegetallen als kenmerken van soorten.

Niche theorie dicteert dat de reactie van soorten op milieugradienten niet lineair is. Elke soort heeft een voorkeursconditie waarbij de soort kan overleven en zich optimaal kan reproduceren. Elke soort komt dus het meeste voor onder een specifiek milieuoptimum en de verdeling van de soort langs een milieugradient is doorgaans unimodaal (eentoppig) met het maximum bij het milieuoptimum. De eenvoudigste unimodale (niet-negatieve) reactie curve voor aan- en

afwezigheidsgegevens is de Gaussisch logistische curve met drie parameters die de niche karakteriseren: optimum (centrum van de niche), tolerantie (breedte van de niche) en maximum (kans van voorkomen in het centrum). Niches van soorten verschillen tussen soorten en van soorten wordt verondersteld dat ze door evolutie aangepast zijn aan het milieu waarin ze leven. Het is moeilijk met de beschikbare software voor (gegeneraliseerde) niet-lineaire gemengde modellen om het Gaussische logistische model met lineaire submodellen voor de parameters, die soortenkenmerken koppelen aan de parameters, aan te passen aan gegevens. In hoofdstuk 3 ontwikkelen we het kenmerk-gemoduleerde Gaussisch logistische model waarin de parameters van de niche lineair afhankelijk zijn gemaakt van de kenmerken van de soorten. Het model wordt aangepast aan gegevens in het Bayesiaanse raamwerk met OpenBugs (Bayesian inference Using Gibbs Sampling). Een Bayesiaanse variabelenselectiemethode wordt gebruikt om die soortskenmerken en milieuvariabelen te identificeren die de aan- en afwezigheidsgegevens het beste verklaren. Ook breiden we de aanpak uit naar het vinden van de beste lineaire combinatie van milieuvariabelen.

In hoofdstuk 4 verklaren we waarom en wanneer gegevens met unimodale structuur effectief kunnen worden geanalyseerd met (gegeneraliseerde) lineair gemengde modellen. We stellen een grafisch hulpmiddel en een statistische toets voor om op unimodaliteit te toetsen. De grafiek en toets maken gebruik van de uitkomsten van een gegeneraliseerd lineair gemengd model zonder enige kwadratische of andere polynomiale term. Een GLMM is natuurlijk een lineair model. Ondanks dit onomstotelijke feit, kan het model gebruikt worden om unimodaliteit te detecteren. Een aanpassing van een GLMM aan gegevens met een unimodale structuur blijkt bruikbaar, als de verschillen in niche breedte tussen soorten niet al te groot is. Als grafisch hulpmiddel stellen we voor om de random plekeffecten uit te zetten tegen de milieuvariabele. Er is aanwijzing voor unimodaliteit, wanneer deze grafiek een kwadratische relatie laat zien. De effectiviteit van GLMM om gegevens met unimodale structuur te analyseren wordt geïllustreerd door de GLMM aanpak te vergelijken met een aanpak met een expliciet unimodaal model voor gesimuleerde en werkelijke gegevens die unimodaliteit laten zien.

Wanneer een systeem wordt beschreven met een statistisch model leidt modelcomplexiteit tot een erg grote rekentijd en een slechte schatting van parameters, in het bijzonder als het aantal voorspellers groot is ten opzichte van de steekproefomvang. Als alternatief voor en verbetering van stapsgewijze methoden zijn krimpmethoden voorgesteld. Eén van deze is de Relevantie vector machine (RVM). RVM kent individuele precisies toe aan de regressiegewichten van voorspellers die dan worden geschat door de marginale aannemelijkheid (likelihood) te maximaliseren (Type II ML of empirisch Bayes). In hoofdstuk 5 onderzoeken we de selectieeigenschappen van RVM, zowel analytisch als experimenteel. We vinden dat RVM nogal tolerant is; het laat voorspellers in

het model die weinig bijdragen. We concluderen dat RVM is geen echte oplossing is voor problemen met hoog-dimensionele gegevens.

In hoofdstuk 6 ontwikkelen we de modelselectiemethode uit hoofdstuk 2 verder, en wel in de context van lineaire gemengde modellen. We noemen de methode gelaagde voorwaartse selectie. In eerste laag, worden random factoren geselecteerd. Dit zijn in dit proefschrift de milieuvariabelen. In de tweede laag worden de vaste effecten geselecteerd (kenmerken van soorten en hun interactie met milieuvariabelen) en in de laatste laag worden de termen die niet significant zijn weer verwijderd. Alle stappen maken gebruikt van een gewijzigd Akaike informatiecriterium. Het lineair gemengde model met gelaagde voorwaartse selectie wordt vergeleken met Type II ML en met bestaande methoden om kenmerk-milieu relaties op te sporen die niet zijn gebaseerd op gemengde modellen, te weten de zogenaamde ‘fourth corner’ (vierde kwadrant) methode en de lineaire kenmerk-milieu methode (LTE).

In hoofdstuk 7 vatten we de bevindingen over de methoden samen. We bespreken de beperkingen van de methoden en geven nieuwe onderzoeksrichtingen aan, zowel voor het modelleren van de verdeling van soorten als voor overeenkomstige uitdagingen in andere disciplines.

Acknowledgements

My PhD has been a wonderful experience for me. The completion of this thesis is a result of hard work, perseverance and the support of several people. I would like to acknowledge those who contributed to it directly or indirectly.

The first and foremost I would like to thank to my supervisor Prof. Cajo ter Braak, who made the most important contribution to my project. It has been a great privilege to work with you. The momentum you created has played a role in the successful completion of this thesis. I have learned so much from you and am thankful for your continuous appreciation, encouragement and support. With your multidimensional approach, you always inspired me. I thank you for putting a lot of faith in me and giving me ample opportunities to explore things. Thank you for being patient and always giving me time when I approached you for scientific guidance/discussion.

I would like to thank Prof. Fred van Eeuwijk, Prof. Jaap Molenaar and Dr Gerie van der Heijden, who always helped me when I needed it. I appreciate your efforts in this regard. Thank you also for financial support for last part of the project.

Wim Ozinga, thank you for your cooperation and data for my first project and the generous help in writing the proposal. Thank you for your positive criticism on my work, it definitely made my work better. Thanks Yu Tong, Carla Kruk and Wout Opdekamp for your professional interaction and providing data.

My journey to PhD was initially started at the Radboud Nijmegen. I am indebted to Prof. Martien Van Zuijlen for what he did for me. It was very unfortunate we couldn't work together because of my personal problems.

Thanks Wies, my office mate, for your company , Thomas, my "R" problem's solver. I warmly thanks Patrica for her friendship and warm company and will never forget her hospitality during my trip to Spain. I also enjoyed being together with Joao, Sabine, Marcos, Santosh, Apri, Yiannis, Alba, Paulo, Laura, Nurudeen, Evert Jan and Martin Boer. Thanks everyone at the Biometris, always willing to help. Thanks Dinie and Hanneke for your support.

Thank you Prof. Dr. Muhammad Aslam for your support during my M.Phil study and also because you were the person who first exposed me to Bayesian statistics.

Thanks Afsheen and Ali bahi for your help in settling down in Netherlands. I cherished your company and the fun with you over the last few years. I wish you all success in the future.

I have enjoyed many things, such as parties, shopping, dinners and body sculpting with Aparna. I appreciate Aparna for her unsuccessful effort to convince me to learn swimming. I am also thankful to Sadia, who took good care of my children.

I am lucky to have a family that has always been a great support. My parents were undoubtedly a guiding force throughout my life and career. Abu and Ami, thank you so much for all your altruistic efforts and prayers. I know, it was very hard for both of you to allow me to go abroad but thank you for believing in me. My gratitude to my mother in-law for her prayers. I am also grateful to my sisters Bushra, Khalida, brother Ali and brother in-laws: Siraj and Rizwan. Thank you so much for your love and cooperation.

My kids deserve a lot of credit and appreciation. They make my life enjoyable and help me to remember that there is so much more to life than just work. Specially, Omamah thank you for keeping Ahsan busy. Meri jaan mama wants to tell you “mama will always be your best friend in the world”. My Ahsan, my curious boy, you underwent a lot. I am gratified to see you now.

A very special thanks for a very special person in my life without whose unconditional faith and support it would have been impossible to achieve this feat. Thank you very much for your ceaseless support, including the final text editing, and epic encouragement during all my PhD. You always kept me motivated. I would also like to thank for sharing all the chores *and* for our travels to the beautiful countries of Europe.

About the Author

Tahira Jamil was born on 12 February, 1973 in Islamabad, Pakistan. She studied Statistics at the University of Punjab, Lahore, Pakistan. After the completion of her MSc, in 1996, she joined the Education Department as Lecturer. Tahira worked as a lecturer at the Govt.Viqar-un-Nisa collage Rawalpindi, where she was involved in teaching statistics to college students. She also supervised/conducted theoretical and practical exams for other academic institutes/Universities. In 2007, Tahira started her PhD at the Biometris, Wageningen University. In 2008, she received a M.Phil degree in Statistics (Allama Iqbad Open University). During her M.Phil thesis entitled "Bayesian Analysis for a Statistical Linear Calibration problem using Noninformative Priors" she got introduce to Bayesian statistics. During her PhD, she also took various statistical courses and skill development courses at Wageningen University and other international universities. She presented her work in national and international conferences. Tahira is married to Muhammad Jamil and proud mother of two children.

List of Publication

- Jamil, T. and C.J.F. ter Braak. Selection properties of Type II maximum likelihood (empirical bayes) in linear models with individual variance components for predictors. Pattern Recognition Letters- *(2011) Submitted*
- Jamil, T., W.A. Ozinga and C.J.F. ter Braak. Selecting traits that explain species-environment relationships: a Generalized Linear Mixed Model approach. Journal of Vegetation Science- *(2011) Submitted*
- Jamil, T. and C.J.F. ter Braak. A Generalized Linear Mixed Model approach to species-trait-environment relationships can handle and detect unimodal relationships with simulated and real data examples- *(2011) Submitted*
- Jamil, T., C. Carla and C.J.F. ter Braak . A unimodal species response model relating traits to environment with application to phytoplankton communities. *(2011) In preparation*
- Jamil, T., W. Opdekamp, van Diggelen, R. and C.J.F. ter Braak. Tiered forward selection in linear mixed models with application to trait selection in a mesocosm experiment- *(2011) In preparation*
- Jamil, M., T. Charnikhova, C. Cardoso, T. Jamil**,** F. Verstappen and H. Bouwmeester. 2011. Quantification of the relationship between strigolactones and Striga hermonthica in rice under varying levels of nitrogen and phosphorus. Weed Research 51:373-385

Thesis

M.Phil thesis: Bayesian Analysis for a Statistical Linear Calibration problem using Noninformative Priors (2008)

Education statement of the Graduate School

PE&RC PhD Education Certificate

With the educational activities listed below the PhD candidate has complied with the educational requirements set by the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC) which comprises of a minimum total of 32 ECTS (= 22 weeks of activities)

Review of literature (4.5 ECTS)

- Models to relate species traits to environment

Writing of project proposal (4 ECTS)

- Bayesian analysis of models to relate species occurrence, specie trait, landscape and environment (2009)

Post-graduate courses (8.2 ECTS)

- Multivariate analysis; PE&RC (2007)
- Applied Bayesian Statistics School on Bayesian methods and econometrics; DEPMQ, Italy (2007)
- Advanced statistics ; PE&RC (2008)
- Mixed models in statistics; CUSO, Lausanne, Switzerland (2009)
- Linear mixed models; PE&RC (2009)
- Bayesian statistics; PE&RC (2009)
- Statistical learning methods for DNA-based of complex traits; PE&RC (2011)

Competence strengthening / skills courses (5.8 ECTS)

- Techniques for writing and presenting a scientific; WGS (2009)
- PhD Competence assessment; WGS (2008)
- Interpersonal communication for PhD students; WGS (2009)
- Scientific writing; WGS (2010)
- Working with Endnote X2; WGS (2010)
- Career perspectives; WGS (2011)

PE&RC Annual meetings, seminars and the PE&RC weekend (1.5 ECTS)

- PE&RC Day (2008)
- PE&RC Weekend (2008)
- PE&RC Day (2009)

Discussion groups / local seminars / other scientific meetings (8.5 ECTS)

- Maths & Stats discussion group (2008-2011)
- Poster presentation: Netherlands Annual Ecology Meeting; Lunteren, the Netherlands (2010)
- Oral presentation: Netherlands Annual Ecology Meeting; Lunteren, the Netherlands (2011)

International symposia, workshops and conferences (3.1 ECTS)

- Mini Symposium Methods for detecting assembly patterns in plant communities; Groningen, the Netherlands (2008)
- Oral presentation: RSS 2010 conference; Brighton, England (2010)

This research was funded by a grant from the Higher Education Commission (HEC), Pakistan through the Netherlands Organisation for International Cooperation in Higher Education (NUFFIC).