Model-based multivariate approaches in community ecology

by

Jonathan Frederik Jupke

Martin-Luther-Straße 41
76829 Landau
Student ID: 212202111
jupk7193@uni-landau.de

SUPERVISION

Prof. Dr. Ralf Schäfer

Dr. Mira Kattwinkel

Master Thesis for the study program MSc. Environmental Sciences
Fachbereich 7: Natur- und Umweltwissenschaften
Universität Koblenz-Landau

7th September 2018

Declaration of Authorship

Hiermit bestätige ich, dass die vorliegende Arbeit von mir selbständig verfasst wurde und ich keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet - Quellen – benutzt habe und die Arbeit von mir vorher nicht in einem anderen Prüfungsverfahren eingereicht wurde.

Jonathan F. Jupke	Date, Location

Acknowledgements

I would like to thank Prof. Schäfer for proposing this thesis topic, helping me through the inevitable road bumps and pushing the simulation process into a productive direction. Further, I owe many thanks to Isabel Müller, Sebastian Scheu, Artem Fischbein and Alisa Bamberg for proofreading the thesis. Their suggestions have provided valuable input. Lastly, I would like to thank Guillaume Blanchet and Matri Anderson, who even though not directly involved in this work, patiently answered my emails.

Contents

1 Introduction

Which environmental gradients drive the changes in species abundances and community composition? This is one of the oldest questions in ecology (e.g. Clements, 1907) and the prospect of humans altering their surroundings at an unprecedented rate endows it with a new urgency (Pacifici et al., 2015). To answer it, ecologists typically record the abundance or occurrence of different taxa and several environmental variables (e.g. precipitation or exposure to stressors), at different sites. This results in a sites-by-species matrix $\mathbf{Y}$ containing multivariate species abundances, which is then statistically related to a sites-by-predictor matrix $\mathbf{X}$, containing the environmental variables. From a statisticians point of view, $\mathbf{Y}$ has many undesirable properties: correlation within and between variables, e.g. through biotic interactions (Morales-Castilla et al., 2015), probability distributions other than the normal, more species than sites (*high dimensionality*, especially in DNA Barcoding studies, Cristescu, 2014) and many zeros, since most species are commonly absent from most sites (*sparsity*).

Multivariate species abundances are frequently analyzed by means of their distance or dissimilarity matrix (distance-based analysis *sensu* Warton et al., 2012). This approach is popular among ecologists because it is non-parametric and hence distribution-free (e.g. Clarke, 1993). Whether a distance metric is appropriate depends on the properties of the data and the aim of the study, as each metric extracts different information from the raw data. The choice is complicated by the vast amount of available metrics (see Legendre and Legendre, 2012) and contradicting recommendations (Faith et al., 1987). Deploying a distance metric, one also implicitly assumes a mean-variance relationship in the data. For example, the Minkowski distances (e.g. Manhattan and Euclidean) assume a constant variance across all mean values (ter Braak and Prentice, 1988). However, species abundances often show a quadratic mean-variance relationship (Routledge and Swartz, 1991; Yamamura, 1999). Miss-specifying the relationship by choosing an improper distance metric can lead to erroneous conclusions about one's data, as was shown by Warton et al. (2012). An alternative to distance-based analysis, that avoids this issue, is the model-based approach.

The model-based approach to multivariate data analysis entails explicitly specifying a statistical model of the process that generated the observed data (Warton et al., 2015b). This includes the mean-variance relationship, which can be adjusted to the properties of the data. Despite their ubiquity in univariate analyses, model-based approaches have long been uncommon in multivariate ecological analyses. However, advances in statistical theory and computation power have led to a surge of models for multivariate abundance data. Recent examples include *Hierarchical Modeling of Species Communities* (HMSC, Ovaskainen et al., 2017), *Generalized Joint Attribute Modelling* (GJAM, Clark et al., 2017) and *multivariate Generalized Linear Models* (GLM$_{mv}$, Warton et al., 2012).

In GLM$_{mv}$, a separate univariate GLM is fit to each taxon, each model using the same predictors.

Univariate GLMs are a powerful and flexible method. They are strongly advocated for the analysis of count or occurence data as they can handle different residual distributions and mean-variance relationships (O'Hara and Kotze, 2010; Warton and Hui, 2011; Szöcs and Schäfer, 2015). Extending them to multi-species abundance data was thus a natural starting point for multivariate model-based analyses (Warton et al., 2012). The univariate models are combined by summing their test statistics, which enables the researcher to draw conclusions about the whole community. GLM_{mv} were one of the earlier multivariate models with an easy-to-use implementation (in the *mvabund* R-package, Wang et al., 2018) and since their introduction they have gained traction within the ecological community (244 citations according to the Web of Science as of 24.04.2018). To my knowledge, the simulations of Warton et al. (2012) remain the only test of GLM_{mv} with simulated data until now. Szöcs et al. (2015) tested GLM_{mv} with data from ecotoxicological mesocosm studies, and found that they performed better or at least as well as commonly used methods (Principal Response Curves). They also emphasized the need for further simulation studies on GLM_{mv}.

The aim of this study is to test the ability of four statistical methods to determine which environmental gradients drive the changes in sets of simulated multivariate abundance data. The performance of GLM_{mv} will be compared to *Distance Based Redundancy Analysis* (db-RDA), *Canonical Correspondence Analysis* (CCA) as well as *Constrained Additive* and *Quadratic Ordination* (CAO/ CQO).
db-RDA is a distance-based analysis. It calculates an ordination on the distance matrix of the sites-by-species matrix $\mathbf{Y}$ constrained by the environmental variables $\mathbf{X}$ (Legendre and Anderson, 1999; Anderson and Willis, 2003). db-RDA was highlighted by Szöcs et al. (2015), because the possibility to use asymmetrical distance metrics and avoid the *double-zero problem* (Legendre and Legendre, 2012) makes them attractive for sparse data sets. CCA and CQO are both solutions to Constrained Gaussian Ordination (CGO), in which species are expected to respond unimodally to latent gradients that are linear combinations of environmental variables (Gauch et al., 1974). CCA is an algorithmic solution (neither distance- nor model-based), that is based on the findings of ter Braak (1986). He showed that the maximum likelihood estimation of CGO for Poisson distributed counts can be approximated by correspondence analysis, given that a set of restrictive assumption hold (see section 2.4). CCA is one of the most widely used statistical techniques in ecology, with the essential papers having accumulated over 3000 citations (ter Braak, 2014). CQO is the maximum likelihood solution to CGO (Yee, 2004). It uses an extension to GLM, *Vector Generalized Linear Models* (VGLM, Yee and Wild, 1996), to estimate model parameters and is hence model-based. Yee (2006) proposed CAO as a modification to CQO, that uses additive instead of linear models and is thus more flexible and data-driven.

2 Material and Methods

2.1 Simulations

Species abundances were simulated as counts, which is the most widely used abundance measure (Warton, 2008b). They responded to an environmental gradient with one of four different response types: unimodal (*uni*), linear (*li*), logistic (*lo*) and bimodal (*bi*). Even though all gradients are linear, a gradient to which species respond unimodally will be called unimodal itself, for the sake of readability.

In the remainder, the following notation is used. $\mathbf{Y}$ is a $N \times S$ matrix of responses, in this case, the abundances of S species, $s = 1 \ldots S$, at N sites, $n = 1 \ldots N$. $\mathbf{X}$ is a $N \times M$ matrix of predictors, here, M environmental variables, $m = 1 \ldots M$.

Unimodal responses were simulated using the Gaussian Response Model (Gauch and Whittaker, 1972a) expanded to multiple gradients (Eqn. 1).

$$y_{s,n} = \prod_{m}^{M_{Uni}} c_{s,m} \times exp\left(- \frac{(x_{m,n} - u_{s,m})^2}{2t_{s,m}^2} \right) \tag{1}$$

In this equation, $u_{s,m}$ is the position of the optimum (i.e. the point with the highest abundance) of species s along the gradient m, $t_{s,m}$ is the tolerance of species s toward the gradient m and determines the width of the curve and $c_{s,m}$ is the maximum abundance of species s on gradient m. M_{uni} is the number of unimodal gradients.
Linear responses were simulated by multiplying the gradient value with a coefficient β (Eqn. 2).

$$y_{s,n} = \prod_{m}^{M_{lin}} x_{m,n} \times \beta_{s,m} \tag{2}$$

Logistic responses were simulated using the Verhulst-equation (Verhulst, 1838, Eqn. 3).

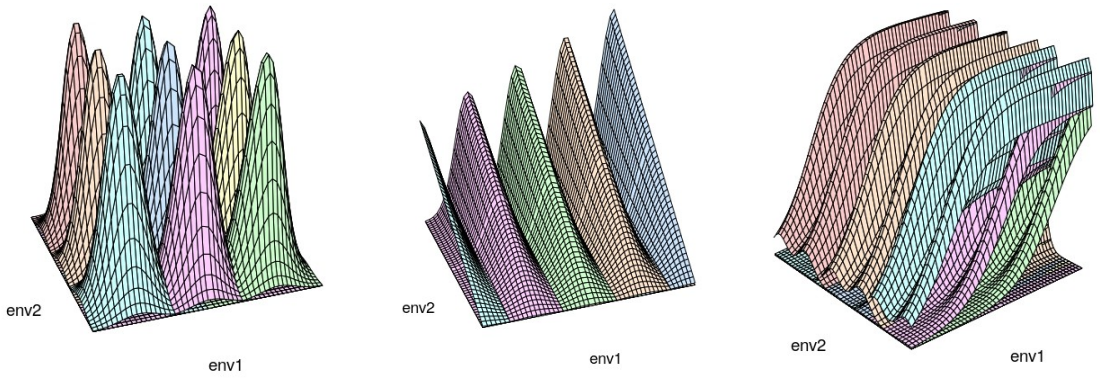
$$y_{s,n} = \prod_{m}^{M_{log}} \frac{c_{s,m}}{1 + exp(-k_{s,m}(x_{0\ s,m} - x_{m,n}))} \tag{3}$$

Here, k determines the steepness of the curve and x_0 is the x-value of the sigmoids midway point.

Bimodal data were simulated by adding two unimodal models with different optima. The total abundance of species s at site n was calculated by multiplying the abundances for individual gradients.

Species were simulated to respond to two environmental gradients: *env1* and *env2*. All species show the same response type towards the same gradient, but response types can differ between gradients. Figure 1 shows three example communities.

This setup allows for ten distinct combinations of response types, including those where the species' response types are the same for both graidents (e.g. Figure 1a). The combinations are referred to by the concatenations of their abbreviated response types, e.g. *Lo-Bi* for a community that responds logistically to the first and bimodally to the second gradient, like the one depicted in Figure 1c.



(a) *Uni-Uni* **(b)** *Uni-Li* **(c)** *Lo-Bi*

Figure 1: Simulated abundance responses along two environmental gradients. Model names are concatenations of abbreviated response types, i.e. (**a**) unimodal-unimodal, (**b**) unimodal-linear and (**c**) logistic-bimodal. The vertical axis indicates abundance as counts. Each example consists of 2500 samples.

The models *Uni-Li*, *Li-Lo* and *Li-Bi* include five species. All other models consist of nine species. They are numbered as follows: for models with nine species, they are numbered first on *env1* from low to high and then on *env2*, i.e. species one's optimum is at a low value for *env1* and *env2* (blue in Figure 1a) and species nine's optimum has a high value in both *env1* and *env2* (pink in Figure 1a). For models with five species, they are numbered from low optimum values to high ones on *env1*. More details on the parameterization of the models are provided in Table 3 in the Supplementary Materials. The two gradients span a grid of 100×100 points for which abundances were simulated. This data set was sampled at equally spaced sampling points and with varying sample sizes (see Table 1). Noise variables were simulated from a standard normal distribution, scaled to the same magnitude as the environmental gradients and restricted to be orthogonal to them and to each other. For each of the ten combinations of response types, four different classes of models were simulated differing in the sample size and number of noise variables. The four classes are shown in Table 1.

Finally, each model was run using three different seeds for random number generation. In total, this gives 120 simulated communities. For a flow chart that visualizes the simulation process see Figure 2.

I used *p*-values to compare the method's performances. While their use has been repeatedly criticized for over 50 years (e.g. Rozeboom, 1960), they provide a valuable means to compare methods as disparate as those tested herein. Three of the four methods (*viz.* GLM_{mv}, db-RDA and CCA) provide *p*-values for the influence of environmental variables on species abundances. Further, the

Table 1: The four model classes. Classes differ in number of samples taken and on the number of noise variables.

Class	Sample size	Noise variables
1	100	1
2	100	5
3	225	5
4	25	5

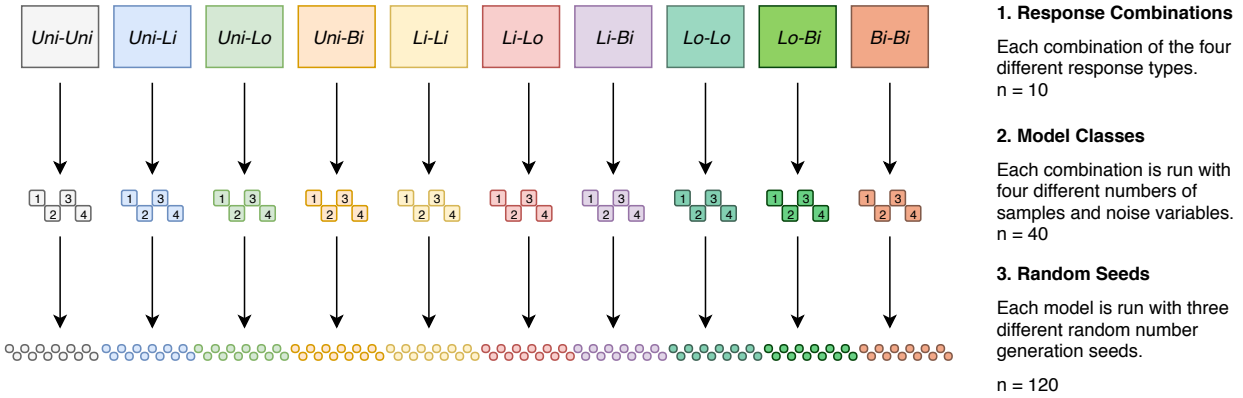


Figure 2: Flowchart of the simulation process. All possible combinations of the four response types unimodal (uni), linear (li), logistic (log) and bimodal (bi) were run in four different classes. The classes differed in how many samples are taken from the simulated community and in how many noise variables are included (see Table 1).

use of *p*-values in ecology and related fields is still wide spread (Fidler et al., 2006) and their performance is thus of practical importance.

2.2 Multivariate Generalized Linear Models

GLM$_{mv}$s are collections of S separately fitted GLMs. Their likelihood ratios for each environmental variable are summed up to obtain the sum-of-Likelihood-Ratios statistic. This test statistic can provide a multivariate *p*-value, showing whether an environmental variable has a statistically significant effect on the mean community abundance. By summing the likelihood ratios one assumes that species responses are independent of each other. This assumption is relaxed in the hypothesis tests because significance is assessed via row resampling, which preserves the correlation structure. Two alternative, computationally more demanding, test statistics are available: the Wald and the score test statistic. Both explicitly account for correlation between variables, by using Generalized Estimation Equations as proposed by Warton (2011).

The effect of each individual species can be assessed by the deviations of the univariate GLMs.

Their p-values are adjusted to multiple testing by controlling the family-wise type I error rate using a resampling-based version of the *Holm's step-down multiple testing procedure* (Westfall and Young, 1993). These adjusted p-values tend to be very conservative (Warton and Popovic, 2018).

Each model was fit with negative binomial, Poisson, and normal residual distribution and their respective canonical links. The fit of the three different models was checked with Dunn-Smith residual-plots and Akaike's Information Criterion (AIC, Akaike, 1974). Hypothesis tests for the best fitting model were conducted using the likelihood ratio statistic, each model was resampled 500 times using the PIT-trap bootstrap procedure (Warton et al., 2017). Besides resampling, no adjustment for inter-species correlations was used.

2.3 Distance Based Redundancy Analysis

db-RDA is the constrained form of Principal Coordinates Analysis (PCoA), an eigenvalue-based ordination conducted on distance matrices. In a PCoA, $\mathbf{Y}$ is transformed into a centered distance matrix $\mathbf{D}$. Eigenvectors of $\mathbf{D}$, are scaled to the length of $\sqrt{\lambda_k}$, with λ_k being the eigenvalue of the k^{th} eigenvector u_k. The scaled eigenvectors are the columns of the principal coordinates matrix $\mathbf{PC}$ (Gower, 1966). The $\mathbf{PC}$ matrix is then used as a response matrix in an RDA. The db-RDA preserves the distance metric which was used to calculate $\mathbf{D}$, which can be metric, semi- or non-metric, setting the method apart from MANOVA (Anderson, 2001) and transformation-based RDA (Legendre and Gallagher, 2001). However, semi- and non-metric distances can produce negative eigenvalues which entail complex ordination axes. These are problematic as they can not be interpreted in a meaningful way. Adding a constant to the squared dissimilarities can correct this (*Lingoes correction*, Gower and Legendre, 1986). Hypothesis tests on the significance of individual constrained axes and environmental variables can be conducted using a pseudo-F-Statistic with permuted residuals (Legendre et al., 2011).
The semi-metric percentage difference distance metric was used. It was introduced by Odum (1950) and is commonly referred to as Bray-Curtis Distance (Legendre and Legendre, 2012). It is asymmetric and thus avoids the double zero problem (Legendre and Legendre, 2012) and is generally acknowledged to have desirable properties for species abundance data (Bloom, 1981; Faith et al., 1987). Hypothesis tests on axes and covariables were conducted with 999 permutations.

2.4 Canonical Correspondence Analysis

CCA has long been the most common way to estimate the parameters of a CGO, even though it is only an approximation of the maximum likelihood solution. It assumes equal tolerances, equal maximal abundances, uniform distribution of species optima and site scores over the latent variable space and bell-shaped responses (ter Braak, 1986). The assumptions are collectively known as the

species packing model. Palmer (1993), Johnson and Altman (1999) and Zuur (1999) confirmed the validity of the approximation and its robustness towards violations against assumptions of the species packing model in simulation studies. Nonetheless, the restrictive assumptions were widely criticized (e.g. Austin and Gaywood, 1994).
An iterative algorithm is used to obtain estimates. First, arbitrary values are assigned to the site scores (positions of sites in latent variable space, $\mathbf{Z}$). These are used to calculate the species optima u (henceforth species scores) as in Eqn. 4.

$$\mathbf{u} = \mathbf{D}_c \mathbf{Y}^t \mathbf{Z} \tag{4}$$

Where $\mathbf{u} = (u_1 \ldots u_S)^t$, $\mathbf{D}_c$ is a diagonal matrix with the abundance of species s across all sites as its s, s^{th} element and $\mathbf{Y}^t$ denotes the transpose of $\mathbf{Y}$.
The species scores are in turn used to calculate the site scores as their weighted average $\mathbf{Z}_{wa}$ (Eqn. 5)

$$\mathbf{Z}_{wa} = \mathbf{D}_r^{-1} \mathbf{Y} \mathbf{u} \tag{5}$$

where $\mathbf{D}_r$ is a diagonal matrix with the abundance of all species at site n as its n, n^{th} element and $\mathbf{D}_r^{-1}$ denotes the inverse of $\mathbf{D}_r$. $\mathbf{Z}_{wa}$ is regressed against $\mathbf{X}$ to obtain the weighted regression coefficient α.

$$\alpha = (\mathbf{X}^t \mathbf{D}_r \mathbf{X})^{-1} \mathbf{X}^t \mathbf{D}_r \mathbf{Z}_{wa} \tag{6}$$

Lastly, $\mathbf{Z}$ is calculated as the product of $\mathbf{X}$ and α. This procedure is repeated until convergence.

The distance between sites (scaling 1) or species (scaling 2) in CCA approximates their two dimensional χ^2-distance, i.e. the Euclidean distance between the expected abundances under the null hypothesis, that abundances do not change along environmental gradients, and the actual data. The absolute difference between expectation and measurement is known as *total inertia*. The ratio of constrained inertia, i.e. variation that can be explained by constrained axes, and total inertia is a general measure of fit for a CCA. The chi-squared-distance is asymmetrical and therefore is unaffected by the double zero problem, but it has been criticized since the same absolute change in a rare species is weighted much stronger than an equal change in an abundant species (Greig-Smith, 1983). The relation between ecological and chi-squared distance is weaker than in other metrics (Faith et al., 1987; Legendre and Gallagher, 2001). The significance tests for axes or environmental variables are calculated using a pseudo-F-Statistic as in db-RDA.

2.5 Constrained Additive and Quadratic Ordination

CQO and CAO are based on VGLMs and their additive counterpart VGAMs, respectively. VGLMs expand GLMs, in that they can encompass multiple response variables, each with a separate linear predictor. The modeled responses can be other parameters than the mean(e.g. the variance) and VGLMs are not restricted to residual distributions from the exponential family.

Both methods employ *Reduced Rank VGLMs/ VGAMs* (RR-VGLM/ VGAM). In RR-VGLMs the M environmental variables are reduced to R latent variables ν. Therefore, the design matrix $\mathbf{X}$ and hat matrix $\mathbf{B}$ are each partitioned into two subsets $\mathbf{X} = (\mathbf{x}_1^T, \mathbf{x}_2^T)^T$; $\mathbf{B} = (\mathbf{B}_1^T, \mathbf{B}_2^T)$. $\mathbf{x}_1$ and $\mathbf{B}_1$ contain covariables and corresponding regression coefficients, which do not contribute to the latent variables. In practice, $\mathbf{x}_1$ often only contains the intercept and $\mathbf{B}_1$ is thus a vector of 1s with length M. $\mathbf{B}_2$ is reduced to a rank R matrix (with full column rank). This is done by reduced rank regression (Anderson, 1951; Izenman, 1975), which determines a low-rank matrix that is an optimal approximation of a full rank matrix. Unlike in db-RDA or CCA, the researcher specifies the number of latent variables, i.e. the rank of $\mathbf{B}_2$, *a priori*. It is referred to as the rank of the method.
The low-rank matrix $\mathbf{B}_2$ is decomposed into two thin matrices $\mathbf{A}$ and $\mathbf{C}$ (see Eqn. 7).

$$\mathbf{B}_2^T = \mathbf{A}\,\mathbf{C}^T \tag{7}$$

The matrix $\mathbf{C}^T$ contains the constrained coefficients, which act as constants in the linear combination of $\mathbf{x}_2$ that constitutes the latent variables. (cf. Eqn.8)

$$\nu = \mathbf{C}^T \mathbf{x}_2 \tag{8}$$

The linear predictor therefore becomes:

$$\eta = \mathbf{B}_1^T \mathbf{x}_1 + \mathbf{A}\nu \tag{9}$$

CQO is an adaption of RR-VGLMs to ecological data sets. It assumes that the response variables show symmetric and bell-shaped responses to the underlying gradients represented by the latent variables. To this end, quadratic RR-VGLM of the kind of Eqn. 10 are used.

$$\eta_s = \beta_{(s)1} x_{(s)1} + \beta_{(s)2}\nu + \beta_{(s)3}\nu^2 \tag{10}$$

The response curve is unimodal for all $\beta_3 < 0$. CQO does not assume the species packing model of the CCA. Three different assumptions can be made concerning the tolerances: (i) equal tolerances, (ii) unequal tolerances and (iii) identity tolerances. The equal tolerance assumption expects tolerance matrices $\mathbf{T}$ to be equal for all species $\mathbf{T}_1 = \mathbf{T}_2 = ... = \mathbf{T}_S$. With the unequal tolerances assumption, the tolerance for each species is estimated separately. For one species $T_s \equiv I_R$, as long as any species has a positive-definite tolerance matrix. Lastly, identity tolerance sets $T_s \equiv I_R$ for all species. This implies the equal tolerance assumption. Equal or identity tolerance models are expected to be faster and easier to interpret but unequal tolerance models should fit the data better (Yee, 2015).
In a CAO the assumptions of symmetric bell shaped responses is relaxed by using a smooth function for ν. The RR-VGAM can be conceptualized as fitting a generalized additive model for each species against the latent variables. As of now, the method is still limited in its capabilities (Yee, 2006). Only rank-1 models can be fitted and only to Poisson or binary responses. Yee (2006) advocates to use CAO for exploratory data analysis and CQO for inference, akin to using Generalized Additive

Models as a diagnostic tool before running GLMs (Hastie et al., 2008).

CAO and CQO were run with Poisson residual distribution and the canonical log-link function. The effective nonlinear degrees of freedom was set to 1.5 as suggested by Yee (2015). Constrained coefficients on the first axis were restricted to be positive. Each model was run ten times and the solution with the lowest deviance was used to safeguard against local solutions. For the calculation of mean values and standard deviations the constrained coefficients of CAO and CQO were transformed to absolute values. To asses the influence of environmental variables on the latent variable, the algebraic sign is not of interest, since it only shows the directionality of the former on the latter. Rank-1 models were run for CAO, rank-1 and rank-2 models with all three tolerance settings for CQO. The optimal number of ranks was found to be two for all models, determined by the AIC as proposed by Yee and Hastie (2003).

2.6 Software

All simulations and analyses were done in R 3.4.4 (R Core Team, 2018). GLM$_{mv}$s were conducted with mvabund 3.13.1. (Wang et al., 2018), db-RDA and CCA with vegan 2.5-2 (Oksanen et al., 2018) and CAO/CQO with VGAM 1.0-5 (Yee, 2018). R-scripts for the simulations as well as the analyses are available on GitHub (https://github.com/JonJup/Master-Thesis)

3 Results

CCA and db-RDA successfully ran for all models, while GLM$_{mv}$s and CAO/CQO had convergence problems. They mostly occurred with class four methods, which have a low sample size to parameter ratio (see Table 1). For GLM$_{mv}$s non-convergence was determined by run time. Models that ran more than 18 hours were aborted and classified as non-converging. Non-convergence of GLM$_{mv}$s occurred for all class four and one class two (Uni-Uni) model. CAO failed to converge for all level four models of *Uni-Li* and *Uni-Bi*; CQO for all class four models of *Uni-Bi*, for unequal and identity tolerance models of *Uni-Uni* and for unequal tolerance models of *Uni-Lo, Lo-Bi* and *Bi-Bi*.

GLM$_{mv}$ was the slowest method. For a class three *Uni-Uni* model the GLM$_{mv}$ ANOVA ran 03:13 h; db-RDA 1.7 min for axis and 4 s for terms; CCA 2 s for axis and $< 1s$ for terms; CAO 9 min and in COQ 28 s for equal tolerance, 9 s for unequal tolerance and 6 s for identity tolerance (on an Ubuntu 18.04 machine with 64-bit, 8 GB RAM and 1.6 GHz).
Means and standard deviations of p-values of GLM$_{mv}$, db-RDA and CCA for all covariables are shown in Table 2.
I refer to the Supplementary Materials for tables of covariable and axes p-values at the level of response combinations or classes (Table 4 - 15), CCA inertias (Table 16 and 17), Canonical

Coefficients of CAO (Table 18 and 19) and CQO (Table 22 - 24) as well as selected ordination diagramms (Figures 9 - 12).

Table 2: *p*-values ± standard deviations of the variables and axes from GLM$_{mv}$, db-RDA and CCA for all models. Not included are CAO and CQO as they do not provide *p*-values. The categories Axes1-3+ do not apply to GLM$_{mv}$ as they do not use latent axes. Axis3+ denotes all axes from the third upwards.

	GLM$_{mv}$	**db-RDA**	**CCA**
env1	0.002 ± 0,001	0.001 ± 0.001	0.201 ± 0.401
env2	0.003 ± 0.002	0.015 ± 0.006	0.101 ± 0.003
Noise	0.579 ± 0.270	0.471 ± 0.257	0.353 ± 0.309
Axis1	NA	0.002 ± 0.003	0.001 ± 0.000
Axis2	NA	0.061 ± 0.176	0.146 ± 0.279
Axis3+	NA	0.604 ± 0.279	0.738 ± 0.374

3.1 Multivariate Generalized Linear Models

In the GLM$_{mv}$s, most models had their lowest AIC and did not violate assumptions with a negative binomial residual distributions. Only in *Li-Li*, *Li-Lo* and *Lo-Lo* was the assumption that residuals are independent of predictors violated, as they showed arched patterns in their residual plots. Based on better fit in the QQ-Plot, a negative binomial distribution was used in *Li-Li* and *Lo-Lo* and a normal distribution in *Li-Lo*.
GLM$_{mv}$'s multivariate *p*-values for both environmental variables, all classes and response type combinations were very low. The *p*-values for noise variables were higher and more spread. The standard deviation of noise variable *p*-values was higher in linear and logistic responses (0.306) than in unimodal and bimodal ones (0.187). The *p*-values of noise variables only fell below the nominal significance level of 0.05 in six models (three form *Li-Bi* and *Lo-Lo* each). For *env1* there was no clear difference between model classes. In *env2*, additional samples decreased the mean *p*-value from 0.003 to 0.002, while also decreasing standard deviation from 0.002 to 0.001. Adding noise variables increased their mean *p*-value from 0.41 to 0.57. Adding samples further increased it to 0.62. (see Table 4 in Supplementary Materials)

Univariate *p*-values were high if a species had its optimum in the middle of an uni- or bimodal gradient (cf. Figure 3). Exceptions were *Lo-Bi* and *Li-Bi*, in which both environmental gradients had low *p*-values for all species. In *Bi-Bi*, *Uni-Li* and *Uni-Lo* the high *p*-values for the first environmental variable were separated into two groups. In all three cases the lower *p*-values were associated with class three models and hence increased sample size. In *Uni-Lo*, species one to three had higher *p*-values than the remaining species for the logistic gradient (see Figure 3). Other than that, both environmental gradients received low *p*-values for all species. Due to the problems with

intermediate uni- or bimodal species, the mean p-values for *env1* and *env2* were high: 0.158 ± 0.34 and 0.134 ± 0.32. When these species are not considered the means drop to 0.003 ± 0.004 and 0.004 ± 0.004.

The mean of noise variable p-values was 0.80 ± 0.24. Adding noise variables and increasing sampling size both lowered the p-values of environmental gradients and increased those of noise variables. The smallest noise variable p-values mostly (25 of 29 below 0.05) originated form class three models.

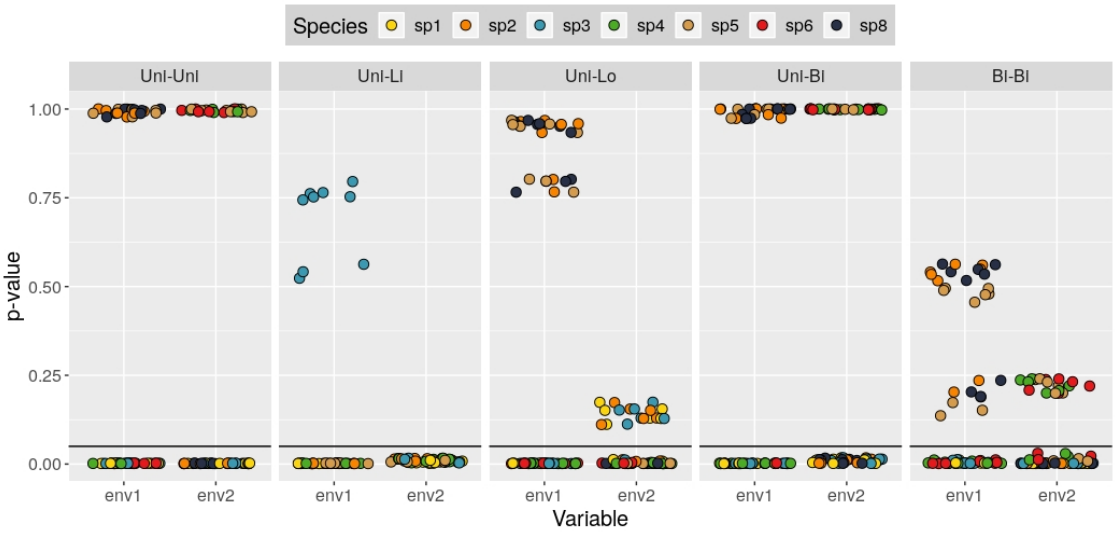


Figure 3: Univariate GLM_{mv} p-values of models containing unimodal or bimodal response types, except *Lo-Bi* and *Li-Bi*. Species that have their optimum at the center of a gradient show higher p-values on the corresponding variable. The black line indicates a p-value of 0.05. Species identity is indicated via point color. Species 7 and 9 are not shown, as their p-values did not show any pattern.

3.2 Distance Based Redundancy Analysis

db-RDA assigned low p-values to most environmental variables. Environmental variables only received high p-values in *Uni-Li* and *Uni-Lo*, where higher p-values were associated with the non-unimodal gradient and class four models. For both environmental variables mean p-values and standard deviations were highest in class four models. The lowest mean p-values for noise variables occurred in *Lo-Lo* at 0.383 ± 0.206. *Bi-Bi* had the most noise variable p-values below 0.05 and was the only model including one in a class one model. Class three models had the lowest mean p-value for noise variables ($Mean \pm SD_{Noise,\ class\ 3} = 0.431 \pm 0.238$) and class four the highest ($Mean \pm SD_{Noise,\ class\ 4} = 0.506 \pm 0.289$).

Mean p-values for the first constrained axis were low, while for the second they were higher and more spread. For all response types and model classes, environmental variables load stronger on the first two constrained axes than the noise variables. The third and higher constrained axes were mostly structured by noise variables and had high p-values. The second axis had higher

p-values in class four models ($Mean \pm SD_{CAP2,\ Class\ 4} = 0.240 \pm 0.288$) than in the other classes ($Mean \pm SD_{CAP2,\ Class\ 1,2,3} = 0.003 \pm 0.002$). This effect is strongest in models with linear or logistic responses.

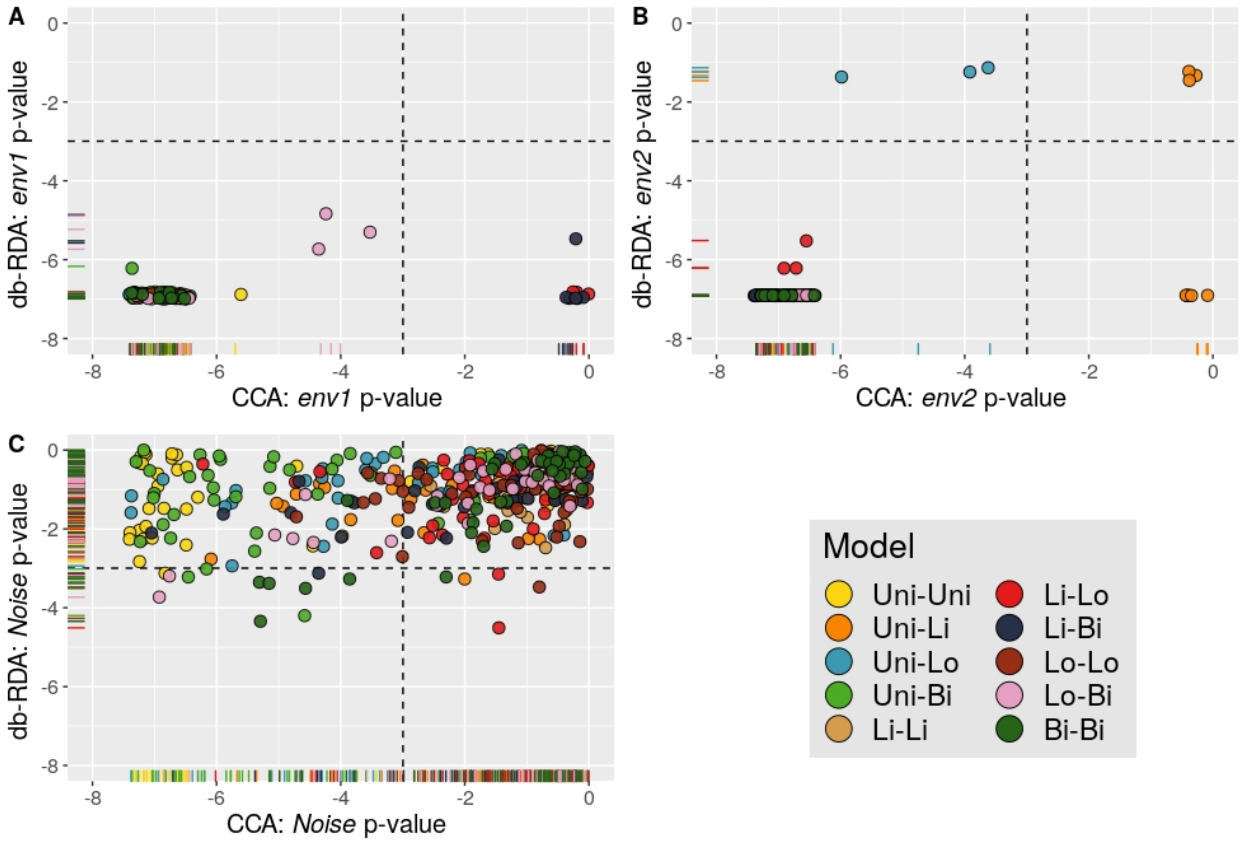


Figure 4: p-values of (**a**) *env1*, (**b**) *env2* and (**c**) *noise variables* for db-RDA and CCA. Y- and X-axis are scaled with a natural logarithm. Black lines indicated p-values of 0.05. Rugs and points are jittered. The color of the points corresponds to the response type combination.

3.3 Canonical Correspondence Analysis

CCA assigned high p-values to linear gradients if only one of the gradients was linear. In *Li-Li*, both gradients received low p-values. Accordingly, mean p-values for *env1* and *env2* are high and wide spread. The p-values for noise variables were low in comparison to the other methods. Leaving out the three models with one linear response (*Uni-Li*, *Li-Lo* and *Li-Bi*), decreased p-values of *env1* and *env2* ($Mean \pm SD_{env1\ \&\ env2} = 0.001 \pm 0.003$), while the mean p-value for noise variables remained at 0.34 ± 0.32.

Noise p-values were especially low in *Uni-Uni* ($Mean \pm SD = 0.155 \pm 0.251$) and *Uni-Bi* ($Mean \pm SD = 0.118 \pm 0.193$). Noise variable p-values in *Bi-Bi* were markedly higher ($Mean \pm SD = 0.507 \pm 0.336$).

The different classes did not have an impact on *env1* and *env2*. In noise variables the p-value

decreased when additional noise variables were added and increased with increasing sample sizes.

The first CCA axis had very low p-values, independent of response type or class ($Mean \pm SD_{CCA1} < 0.001 \pm 0$). For the response combinations *Uni-Li*, *Li-Lo* and *Li-Bi* the second axis had high p-values ($Mean \pm SD = 0.459 \pm 0.336$), but was also strongly determined by noise variables. Overall, the mean p-value for the second axis was higher than for the first. Removing the three models *Uni-Li, Li-Lo* and *Li-Bi* reduced its p-value to 0.011 ± 0.055. In *Lo-Bi*, the second axis had considerably higher p-values in the class four models ($Mean \pm SD_{Lo-Bi,\ class4} = 0.260 \pm 0.170$) than in the other three classes ($Mean \pm SD_{Lo-Bi,\ class\ 1,2,3} = 0.001 \pm < 0.001$). In all response types, the axes three and higher were determined strongest by noise variables and p-values were high. *Uni-Uni* and *Uni-Bi* had low p-values for the third and higher axes: $Mean \pm SD_{CCA3+,Uni-Uni} = 0.364 \pm 0.416$; $Mean \pm SD_{CCA3+,Uni-Bi} = 0.407 \pm 0.407$.

Total inertia differed between response types: Uni- and bimodal responses had the highest total inertia ($Mean \pm SD_{Uni/Bi-X} = 4.45 \pm 1.94$), linear and logistic the lowest ($Mean \pm SD_{Li/Lo-X} = 0.15 \pm 0.13$). Mixtures of these response types lay in between ($Mean \pm SD_{Li/Lo-Uni/Bi} = 1.60 \pm 0.80$). This pattern was also found in constrained and unconstrained inertia (see Supplementary Materials Table 13).

3.4 Constrained Additive and Quadratic Ordination

For models with only linear or logistic responses, i.e. *Li-Li, Li-Lo* and *Lo-Lo*, CAO failed to estimate any species parameters. Its also in these models and in *Bi-Bi*, that both variables contributed equally to the latent variable. In the others, one variable determined the latent gradient while the other contributed as much as the noise variables (see Figure 5). *Uni-Lo* was the only response combination for which all parameters were estimated in all classes. In all other combinations, CAO failed to estimate at least some parameters. In *Uni-Bi* and *Uni-Uni* estimation failed only for class four models.
Species maxima were underestimated for every response combination. The mean difference between estimated and actual maxima, expressed in percent of actual maxima was $-84.9\ \% \pm 9.3\ \%$.

Constrained coefficients did not differ significantly between tolerance types, but varied strongly among response types (see Figure 6). The weights of uni- or bimodal responses were higher than those of logistic or linear ones. The bimodal gradient in class four, unequal tolerance *Li-Bi* (0.277) had the highest coefficient and the linear gradient in the same model the lowest (0.001). The weight of a gradient depended not only on the response type but also on the other gradient. For example, the mean coefficient of the unimodal gradient in *Uni-Uni* is 0.18, while it is 0.09 in *Uni-Li*. The mean weights for *env1* and *env2* were 0.079 ± 0.062 and 0.064 ± 0.049. Outliers are often results of class four models, for example, in *Uni-Uni* and *Li-Bi*. However, in some models (e.g. *Uni-Bi*)

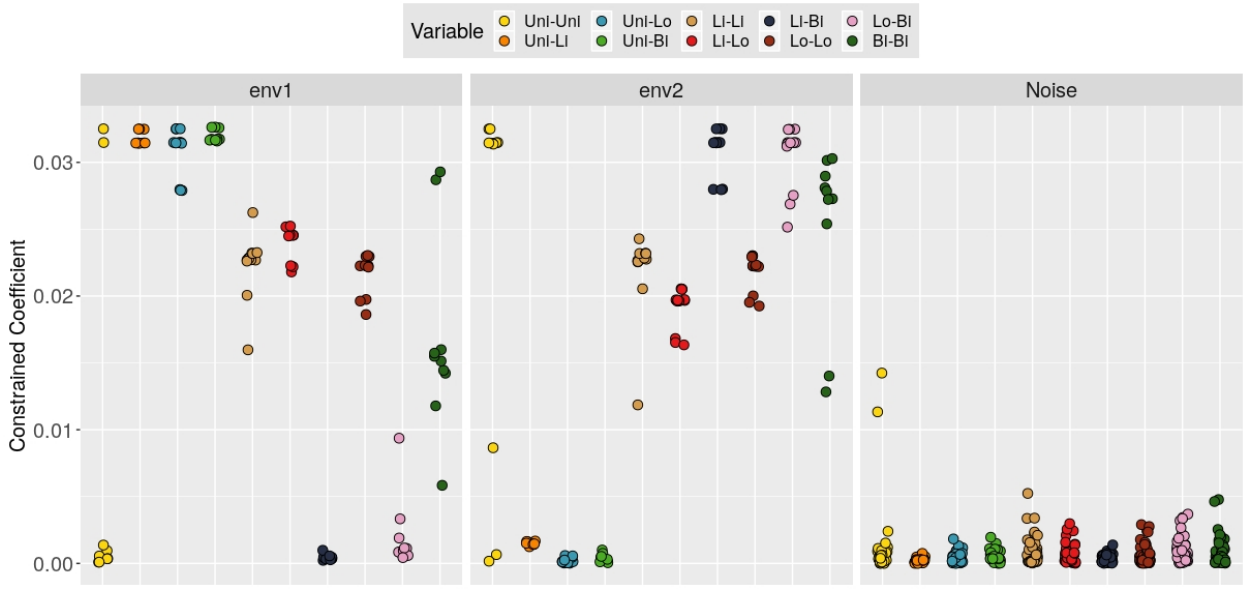


Figure 5: Absolute values of constrained coefficients from the Constrained Additive Ordination. Colors indicate the response type combination. Each box contains the values for one covariable. Points are horizontally jittered.

outliers were generated by other model classes.
Overall, the weights of noise variables were markedly lower than those of environmental variables ($Mean \pm SD_{Noise} = 0.002 \pm 0.003$). The coefficient was highest in *Uni-Uni* with 0.006 and lowest in *Lo-Lo* with 0.001. More noise variables and fewer samples increased the weights of noise variables and decreased those of environmental variables. Additional samples had the opposite effect.

There was no overarching effect of either class or tolerance assumption on constrained coefficients. The only notable effect was the lower convergence rate of unequal tolerance models. The estimated maxima were extremely inaccurate. In most cases, they were too high, but occasionally also too small. An illustrative example is the equal tolerance, class three, *Uni-Bi* model: For two seeds the maximum abundance of species 9 was overestimated by 1.7e+98 % of the actual maximum and for the other seed it was underestimated by 100 %.

4 Discussion

I ran four different statistical methods (GLM_{mv}, db-RDA, CCA, CAO/CQO) on a total of 120 simulated abundance datasets, that differed in response types, sample sizes, number of noise variables and random number generation seeds, to assess their ability to differentiate between environmental variables and noise variables.

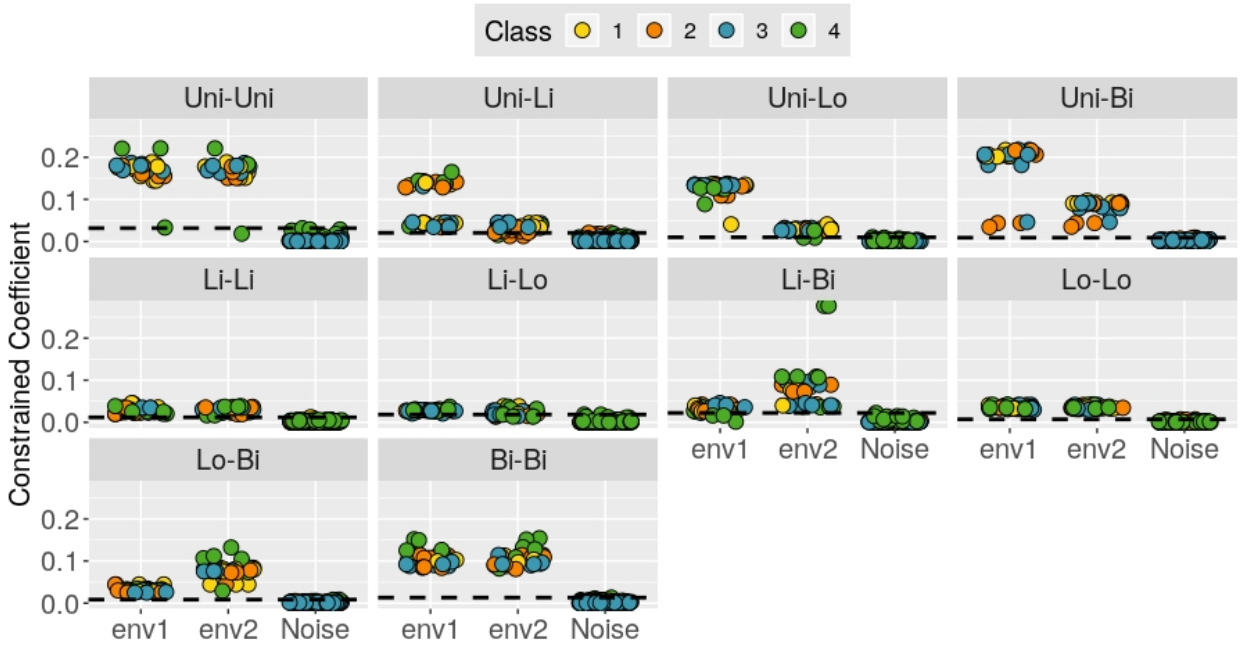


Figure 6: Constrained Coefficients of explanatory variables in a Constrained Quadratic Ordination. Each box is one response type combination. Point colors indicate the model class. The black line marks the highest weight of a noise variable for that response type combination. The points are horizontally jittered.

The GLM$_{mv}$ correctly assigned low multivariate p-values to all environmental variables, and higher ones to noise variables. Univariate significance tests also performed well, except for uni- or bimodal gradients receiving high p-values for species with optima close to their centers. For class four models GLM$_{mv}$s did not converge. db-RDA also differentiated well between environmental and noise variables for all response type combinations. However, low sample sizes increased p-values for environmental variables and decreased those of noise variables. In CCA, linear gradients got high p-values if combined with other response types and noise variables had the lowest p-values out of all methods used. CAO and CQO showed convergence problems, especially in unequal tolerance CQOs. CAO only estimated all parameters in one of the models (*Uni-Lo*) and failed to estimate any parameters in all models with only linear and logistic gradients. As was to be expected, CQO performed best with uni- and bimodal gradients. The constrained coefficient of linear and logistic gradients were frequently at the level of noise variables.

GLM$_{mv}$s were least influenced by different response types. They were, however, affected strongest by low sample sizes, as all class four models failed to converge. The lowest multivariate p-values for noise variables occurred in models that violated the random residuals assumptions (*Li-Bi* and *Lo-Lo*) and thus would likely not be regarded as reliable. Species have high univariate p-values for a gradient if they reach their optimum near or at the middle of the gradient's sampled range. This is caused by low estimated regression coefficients, which is likely due to the symmetrical nature of the response (see Response Symmetry in GLMs in the Supplementary Materials). As it is highly un-

likely to observe perfectly symmetrical response shapes right at the middle of the sampled gradient range, it is fair to assume that these problems were due to the simulation set up and improbable in field data. Another problem of GLM$_{mv}$ is the long runtime, which is due to the resampling (Wang et al., 2012). By resampling, one avoids having to specify the correlation between species in the model. The correlation is accounted for at the inference stage by resampling observations across independent sites (Anderson, 2001). Models that include the correlation structure explicitly avoid resampling and thus can reduce computation time. First advances in this direction have been made, e.g. by Jamil et al. (2012) who used the site effect of a Generalized Linear Mixed Model (GLMM) to induce equal correlation between all species pairs.
In contrast to the other considered methods, GLM$_{mv}$s do not reduce the dimensions of the data. Visualizing the multivariate and thus multidimensional data is therefore cumbersome. Indeed, no easy-to-use and to interpret method to present the output of GLM$_{mv}$s is available.
Many features of GLM$_{mv}$s remain unexplored in this study. I only tested the model under the assumption of independence of species. Other correlation type settings can account for dependence between species, either with an unstructured correlation matrix (only advisable when $N > S$) or by shrinking the correlation matrix towards identity using ridge regularization (Warton, 2008a). These correlation types necessitate the usage of another test statistic than the likelihood ratio. Currently, the Wald test statistic and the score statistic are also implemented to this end. The Wald test statistic makes use of GEE with the sandwich-type-estimator of Warton (2011) but is unsuitable for count or occurrence data with means at zero. For such cases, the score statistic should be used. Datasets simulated to test these variants of GLM$_{mv}$ could highlight performance differences even more strongly, as the other methods do not incorporate adjustments to these properties.

db-RDA also proved to be robust to different response types and sample sizes. Although, low samples sizes decreased the method's power in *Uni-Lo* and *Uni-Li*. The six models with high *env2* p-values (cf. Figure 4) are all class four models. db-RDA's good performance is in concert with other simulation studies (e.g. Roberts, 2009). Hence, its popularity in the ecological community seems justified. These results are only valid for the Bray-Curtis Distance metric, which was used here. Other measures would likely have produced different results, therefore the selection of an appropriate metric is a crucial step in any db-RDA analysis. Having to choose a single metric can be avoided by using consensus RDA (Blanchet et al., 2014). In this novel method, multiple db-RDA are run, only differing in their distance metric. Site scores on statistically significant axes are combined into one matrix which acts as response matrix in a new RDA. This method extracts the information that is common to all individual models. Simulation studies comparing properties of consensus RDA with those individual db-RDA and other methods, distance- or model-based, are lacking. Another avenue for future development of db-RDA would be novel distance metrics, but there have been no recent developments (M. J.Anderson, pers. comm.).

The CCA performed worst of the methods tested. Linear responses were categorically assigned

high *p*-values if they were combined with other response types. The method should thus not be used for combinations of linear and non-linear responses. Linear response usually occur if the sampled gradient is short relative to the species' tolerance. For axes as well as terms the noise *p*-values were lower than in other methods. Most low *p*-values for noise variables in CCA occurred in uni- or bimodal models. Given that *Uni-Uni* fits the expectations of the species packing model and bimodal models do not deviate strongly (responses are symmetric but not bell-shaped), this is surprising. These models also had higher inertia than those with linear or logistic gradients, which suggests that the type-I-error rate might be related to the inertia (see Supplementary Materials Table 16 and 17). Given that the triplots (see Supplementary Materials Figure 10) are not skewed by noise variables with low *p*-values, the problem might also reside in the method that is used to calculate the *p*-values. Newer approaches to CCA that can correct for zero inflation (Zhang and Thas, 2012) or non-linear relationships between predictor and response variable (Makarenkov and Legendre, 2002) are available but not widely used. ter Braak and Šmilauer (2015) advocate to move beyond Gaussian Ordination and further develop methods that are based on smooth functions, like the CAO.

In CAO, estimation of optima and maxima failed in all models that only included linear and logistic responses. In most other models, estimates for some species were missing, particularly those with optima close to the end of the sampling range. Species maxima were underestimated in all response type combinations. This was strongest in *Uni-Uni* and *Uni-Bi*, which adhere the closest to the assumptions of CQO.
In most models, only one gradient significantly influenced the latent variable (see Figure 5), even if both gradients were identical (e.g. *Uni-Uni*). This problem might improve with rank-2 CAOs but further tests, also with more than two environmental variables, are necessary. Using a CAO to test for unimodality and thus whether to use or to discard a variable for a CQO was rather conservative for these data. Parameter estimation even failed for some species in *Uni-Uni* and *Uni-Bi*. Jamil and ter Braak (2013) present an alternative, graphical tool based on GLMMs to test for unimodality, which was applied by Jamil et al. (2014), though not in connection with CQO. Overall, the CAO was no reliable guide CQO-suitability and estimated species parameters systematically diverged from the actual values. However, many features of this method, e.g. rank-2 models and further residual distributions, are not yet implemented in VGAM (Yee, 2015). Alternative methods using smooth functions for ordination are also available (e.g. Schnabel et al., 2012)

The weight that CQO assigned to different response types varied strongly. The mean weight of environmental gradients was always higher than that of noise variables, but some individual linear or logistic gradients have constrained coefficients on the level of noise variables. Uni- and bimodal gradients had higher weights than linear or logistic ones. However, the response types also influenced each other. The weight of uni- or bimodal gradients was much lower when combined with linear or logistic gradients instead of each other. How the influence of gradient types on each others constrained coefficient develops in settings with more than just two variables, as one would

expect in actual communities, would be an interesting avenue for further studies.
The impact of different numbers of samples and noise variables was, besides the convergence problems with some class four models, negligible. The same holds true for different tolerance settings. The most notable difference was the lower convergence rate in unequal tolerance models. Equal tolerance models ran longer than those with unequal and identity tolerance. Since tolerances have to be estimated individually in unequal tolerance models these were expected to run the longest (Yee, 2015). The estimates for abundance maxima were very inaccurate and mostly too high. Based on the strength of divergence between estimated and real values, I advise against using these maxima.
Up until now, usage of CQO and CAO is seldom in ecological studies and mostly within fisheries research (e.g. Vilizzi et al. (2012), Top et al. (2016) and Carosi et al. (2017)). ter Braak and Šmilauer (2015) suggest that this is due to limitations on the number of species that can be included, a steep learning curve and numerical instability. This study confirmed that in its current state the method has issues with convergence and parameter estimations. It is especially concerning that the weights of unimodal gradients can be decreased by other non-unimodal gradients and that estimated maxima are overestimated.

Further steps can be taken, to render the simulated data more alike actual field data. Austin (1999) criticized unimodal and symmetric response curves as overly simplistic and Austin et al. (1994) proposed beta-functions to simulate unimodal but asymmetrical shapes. However, in a study of Oksanen and Minchin (2002) only about 20% of the responses were strongly skewed, while symmetric and bell shaped responses were the most common. In this study, I opted to use bimodal and logistic responses as skewed versions of the bell-shape instead of beta functions.
Also, species abundances were only determined by the environmental gradients. Stochasticity can be added to one or to both ends of the relationship. For example, in form of a random term, which is added to the abundances or environmental variables (e.g McCune, 1997). Likewise, observation errors can be included via binomial functions, as is done in N-mixture models (Royle, 2004), to examine how susceptible a method is towards regression dilution (e.g. Frost and Thompson, 2000; McInerny and Purves, 2011). Lastly, when noise is added to response and explanatory variables this would, given the noise is simulated from the same distribution both times, likely result in endogeneity, i.e. a non-zero covariance between the residuals and one or more explanatory variables. Simulations with induced endogeneity would be interesting as this phenomenon might be underappreciated by ecologists (Armsworth et al., 2009; Fox et al., 2015).

The methods that were compared in this study have never been directly compared before. Most similar is the work of Warton et al. (2012), who compare GLM_{mv} to CCA and RDA (not db-RDA) among others. They showed that only GLM_{mv} successfully differentiates between location effect (difference in means) and dispersion effect (difference in variance). However, comparative studies of multivariate methods, in general, are common. Especially ordination techniques like CCA and

RDA were subject to extensive testing in the 1970s and 1980s (e.g. Gauch and Whittaker, 1972b; Gauch et al., 1977; Kenkel and Orloci, 1986). To my knowledge, Roberts (2008) and Roberts (2009) are the only studies that methodically compare db-RDA to other methods. Both compare db-RDA, CCA and Multidimensional Fuzzy Set Ordinations (MFSO). Roberts (2008) uses simulated data sets to this end, while Roberts (2009) uses four different sets of field data. Both studies concur that db-RDA outperforms CCA and is tied with or performs slightly inferior to MFSO. CQO/CAO are occasionally tested in comparisons of individual and community level species distribution models (e.g. Baselga and Araújo (2009) and Maguire et al. (2016)), where they are an instance of the latter. They were generally not found to significantly improve upon the individual models (e.g. GLMs or Regression Trees).

In this study, one (GLM_{mv}) model-based approach outperformed both the distance-based and the algorithmic approach, while the other (CQO) was not clearly superior, at least not to db-RDA. Both, GLM_{mv} and db-RDA proved robust towards any response type tested herein, while both having poorer properties with little sample sizes. CCA should not be used for linear responses, i.e. short gradients, and the term p-values are very low for noise variables, especially in models with high inertia. The comparison to CAO and CQO is complicated by the absence of a common statistic. Both showed issues with convergence and estimation accuracy. Constrained Coefficients varied strongly between response types and within some models (CAO e.g. *Uni-Uni*, CQO e.g. *Li-Bi*). Even for models with unimodal response type CQO severely overestimated the maximal abundances, this parameter should hence not be used. This study thus showed, that some but not all approaches in model-based multivariate inference have considerable potential and can outperform more common distance-based or algorithmic methods. As these models are still at an early stage, new developments and increases in computation speed can be expected.

An especially active area are models using joint probability distributions (e.g. Clark et al., 2014; Pollock et al., 2014). While GLM_{mv} estimate the mean of the conditional distribution Y_s given $\mathbf{X}$ ($\hat{E}(Y_s|\mathbf{X})$) separately for each species and CQO combine all species to estimate the latent variable but regress each species separately against them, joint models estimate the joint distribution of all species conditional on the environmental variables. A common interest of many joint models is to imply biotic interactions from the residuals of the species-environment interaction, as these two sets of predictors (biotic and biotic) were shown to have little redundancy (Meier et al., 2010). Some of the models also anticipate the challenges of Big Data for ecology (Hampton et al., 2013). *Generalized Linear Latent Variable Models*, for example, include latent variables instead of random effects to capture residual correlation, which significantly reduces the size of the variance – covariance matrix (Warton et al., 2015a; Niku et al., 2017). In HMSC (Ovaskainen et al., 2017) this approach is coupled with a fourth corner model (including species traits, Legendre et al., 1997) and phylogenetic relationships to make sense of many types of data. To the same end, GJAMs allow for different kinds of data (e.g. continuous, discrete counts, ordinal counts and occurrence) to be

included in the same response variable and have outperformed Poisson GLM on discrete count data and a Bernoulli GLM on binary host status data in a recent simulation study (Clark et al., 2017).

It is now essential that ways to infer ecological processes from the modeled patterns develop at a similar pace, to avoid confusing statistical artifacts with genuine biological signals (Dormann et al., 2018). If this succeeds, a move form distance-based and algorithmic methods towards model-based methods might entail one from the current implicit Gleassonian towards a from of modern Clementsian perspective (Eliot, 2011); from asking how do individual species change along environmental gradients towards asking how do communities change as a whole.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Australian Ecology*, 26(1):32–46.

Anderson, M. J. and Willis, T. J. (2003). Canonical Analysis of Principal Coordinates: A Useful Method of Constrained Ordination for Ecology. *Ecology*, 84(2):511–525.

Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics*, 22(3):327–351.

Armsworth, P. R., Gaston, K. J., Hanley, N., and Ruffell, R. (2009). Contrasting approaches to statistical regression in ecology and economics. *Journal of Applied Ecology*, 46(2):265–268.

Austin, M., Nicholls, A., Doherty, M., and Meyers, J. (1994). Determining species response functions to an environmental gradient by means of a β-function. *Journal of Vegetation Science*, 5(2):215–228.

Austin, M. P. (1999). A Silent Clash of Paradigms: Some Inconsistencies in Community Ecology. *Oikos*, 86(1):170 – 178.

Austin, M. P. and Gaywood, M. J. (1994). Current problems of environmental gradients and species response curves in relation to continuum theory. *Journal of Vegetation Science*, 5(4):473–482.

Baselga, A. and Araújo, M. B. (2009). Individualistic vs community modelling of species distributions under climate change. *Ecography*, 32(1):55–65.

Blanchet, F. G., Legendre, P., Bergeron, J. A. C. B., and He, F. (2014). Consensus RDA across dissimilarity coefficients for canonical ordination of community composition data. *Ecological Monographs*, 84(3):491–511.

Bloom, S. A. (1981). Similarity indices in community studies: potential pitfalls. *Marine Ecological Progress Series*, 5(2):125–128.

Carosi, A., Ghetti, L., La Porta, G., and Lorenzoni, M. (2017). Ecological effects of the European barbel Barbus barbus (L., 1758) (Cyprinidae) invasion on native barbel populations in the Tiber River basin (Italy). *European Zoological Journal*, 84(1):420–435.

Clark, J. S., Gelfand, A. E., Woodall, C. W., and Zhu, K. (2014). More than the sum of the parts: Forest climate response from joint species distribution models. *Ecological Applications*, 24(5):990–999.

Clark, J. S., Nemergut, D., Seyednasrollah, B., Turner, P. J., and Zhang, S. (2017). Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data. *Ecological Monographs*, 87(1):34–56.

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, 18(1):117–143.

Clements, F. E. (1907). *Plant Physiology and Ecology*. New York, NY: Henry Holt and Company.

Cristescu, M. E. (2014). From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *Trends in Ecology & Evolution*, 29(10):566–571.

Dormann, C., Bobrowski, M., Dehling, M., Harris, D., Hartig, F., Lischke, H., Moretti, M., Pagel, J., Pinkert, S., Schleuning, M., et al. (2018). Biotic interactions in species distribution modelling: Ten questions to guide interpretation and avoid false conclusions. *Global Ecological Biogeography*, 00:1–13.

Eliot, C. (2011). The legend of order and chaos: communities and early community ecology. *In: Brown, B., de Laplante, K. and Peacock, K.: Philosophy of Ecology. Elsevier, Netherlands.*, pages 49–107.

Faith, D. P., Minchin, P. R., and Belbin, L. (1987). Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio*, 69(1-3):57–68.

Fidler, F., Burgman, M. A., Cumming, G., Buttrose, R., and Thomason, N. (2006). Impact of Criticism of Null-Hypothesis Significance Testing on Statistical Reporting Practices in Conservation Biology. *Conservation Biology*, 20:1539–1544.

Fox, G. A., Negrete-Yankelevich, S., and Sosa, V. J. (2015). *Ecological statistics: contemporary theory and application*. Oxford University Press, USA.

Frost, C. and Thompson, S. G. (2000). Correcting for regression dilution bias: comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(2):173–189.

Gauch, H. G. J., Chase, G. B., and Whittaker, R. H. (1974). Ordination of vegetation samples by gaussian species distributions. *Ecology*, 55(6):1382–1390.

Gauch, H. G. J. and Whittaker, R. H. (1972a). Coenocline Simulation. *Ecology*, 53(3):446–451.

Gauch, H. G. J. and Whittaker, R. H. (1972b). Comparison of ordination techniques. *Ecology*, 53(5):868–875.

Gauch, H. G. J., Whittaker, R. H., and Wentworth, T. R. (1977). A comparative study of reciprocal averaging and other ordintion techniques. *Journal of Ecology*, 65(1):157–174.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338.

Gower, J. C. and Legendre, P. (1986). Metric and euclidean properties of dissimilarity coefficients. *Journal of classification*, 3(1):5–48.

Greig-Smith, P. (1983). *Quantitative Plant Ecology*, volume 9. University of California Press, USA.

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., Duke, C. S., and Porter, J. H. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment*, 11(3):156–162.

Hastie, T., Tibshirani, R., and Friedman, . J. (2008). *The Elements of Statistical Learning*. Springer, New York., 2nd edition.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264.

Jamil, T., Kruk, C., and ter Braak, C. J. (2014). A unimodal species response model relating traits to environment with application to phytoplankton communities. *PloS one*, 9(5):e97583.

Jamil, T., Ozinga, W. A., Kleyer, M., and ter Braak, C. J. F. (2012). Selecting traits that explain species environment relationships : a Generalized Linear Mixed Model approach. *Journal of Vegetation Science*, 24:1–43.

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

Johnson, K. W. and Altman, N. S. (1999). Canonical Correspondence Analysis as an approximation to Gaussian ordination. *Environmetrics*, 10(1):39–52.

Kenkel, N. C. and Orloci, L. (1986). Applying Metric and Nonmetric Multidimensional Scaling to Ecological Studies : Some New Results. *Ecology*, 67(4):919–928.

Legendre, P. and Anderson, M. J. (1999). Distance-Based Redundancy Analysis: Testing Multispecies Responses in Multifactorial Ecological Experiments. *Ecological Monographs*, 69(1):1–24.

Legendre, P. and Gallagher, E. D. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia*, 129(2):271–280.

Legendre, P., Galzin, R., and Harmelin-Vivien, M. L. (1997). Relating behavior to habitat: Solutions to the fourth-corner problem. *Ecology*, 78(2):547–562.

Legendre, P. and Legendre, L. F. J. (2012). *Numerical Ecology*, volume 24 of *Developments in Environmental Modelling*. Elsevier, 3 edition.

Legendre, P., Oksanen, J., and ter Braak, C. J. F. (2011). Testing the significance of canonical axes in redundancy analysis. *Methods in Ecology and Evolution*, 2(3):269–277.

Maguire, K. C., Nieto-Lugilde, D., Blois, J. L., Fitzpatrick, M. C., Williams, J. W., Ferrier, S., and Lorenz, D. J. (2016). Controlled comparison of species- and community-level models across novel climates and communities. *Proceedings of the Royal Philosophical Society - B*, 283:20152817.

Makarenkov, V. and Legendre, P. (2002). Nonlinear redundancy analysis and canonical correspondence analysis based on polynomial regression. *Ecology*, 83(4):1146–1161.

McCune, B. (1997). Influence of noisy environmental data on canonical correspondence analysis. *Ecology*, 78(8):2617–2623.

McInerny, G. J. and Purves, D. W. (2011). Fine-scale environmental variation in species distribution modelling: regression dilution, latent variables and neighbourly advice. *Methods in Ecology and Evolution*, 2:248–257.

Meier, E. S., Kienast, F., Pearman, P. B., Svenning, J. C., Thuiller, W., Araújo, M. B., Guisan, A., and Zimmermann, N. E. (2010). Biotic and abiotic variables show little redundancy in explaining tree species distributions. *Ecography*, 33(6):1038–1048.

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

Niku, J., Warton, D. I., Hui, F. K., and Taskinen, S. (2017). Generalized Linear Latent Variable Models for Multivariate Count and Biomass Data in Ecology. *Journal of Agricultural, Biological, and Environmental Statistics*, 22(4):1–25.

Odum, E. P. (1950). Bird populations of the highlands (north carolina) plateau in relation to plant succession and avian invasion. *Ecology*, 31(4):587–605.

O'Hara, R. B. and Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, 1(2):118–122.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., and Wagner, H. (2018). *vegan: Community Ecology Package*. R package version 2.4-6.

Oksanen, J. and Minchin, P. R. (2002). Continuum theory revisited: What shape are species responses along ecological gradients? *Ecological Modelling*, 157(2-3):119–129.

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T., and Abrego, N. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters*, 20(5):561–576.

Pacifici, M., Foden, W. B., Visconti, P., Watson, J. E., Butchart, S. H., Kovacs, K. M., Scheffers, B. R., Hole, D. G., Martin, T. G., Akcakaya, H. R., et al. (2015). Assessing species vulnerability to climate change. *Nature Climate Change*, 5(3):215.

Palmer, M. W. (1993). Putting Things in Even Better Order: The Advantages of Canonical Correspondence Analysis. *Ecology*, 74(8):2215–2230.

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

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Roberts, D. W. (2008). Statistical Analysis of Multidimensional Fuzzy Set Ordinations. *Ecology*, 89(5):1246–1260.

Roberts, D. W. (2009). Comparison of multidimensional fuzzy set ordination with CCA and DB-RDA. *Ecology*, 90(9):2622–2634.

Routledge, R. D. and Swartz, T. B. (1991). Taylor's power law re-examined. *Oikos*, 60(1):107–112.

Royle, J. A. (2004). N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, 60(1):108–115.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, 57(5):416.

Schnabel, S. K., van Eeuwijk, F. A., and Eilers, P. H. (2012). Modeling latent curves for genotype by environment interaction. In *Proceedings of the 27th International Workshop on Statistical Modellin Prague*, pages 309–313.

Szöcs, E. and Schäfer, R. B. (2015). Ecotoxicology is not normal. *Environmental Science and Pollution Research*, 22(18):1399013999.

Szöcs, E., Van den Brink, P. J., Lagadic, L., Caquet, T., Roucaute, M., Auber, A., Bayona, Y., Liess, M., Ebke, P., Ippolito, A., ter Braak, C. J., Brock, T. C., and Schäfer, R. B. (2015). Analysing chemical-induced changes in macroinvertebrate communities in aquatic mesocosm experiments: a comparison of methods. *Ecotoxicology*, 24(4):760–769.

ter Braak, C. J. (2014). History of canonical correspondence analysis. *In Blasius J. and Greenacre M. Visualization and Verbalization of Data. CRC Press, Boca Raton, USA*, pages 61–75.

ter Braak, C. J. and Šmilauer, P. (2015). Topics in constrained and unconstrained ordination. *Plant Ecology*, 216(5):683–696.

ter Braak, C. J. F. (1986). Canonical Correspondence Analysis: A New Eigenvector Technique for Multivariate Direct Gradient Analysis. *Ecology*, 67(5):1167–1179.

ter Braak, C. J. F. and Prentice, I. C. (1988). A Theory of Gradient Analysis. *Advances in Ecological Research*, 18:271–317.

Top, N., Tarkan, A. S., Vilizzi, L., and Karaku, U. (2016). Microhabitat interactions of non-native pumpkinseed Lepomis gibbosus in a Mediterranean-type stream suggest no evidence for impact on endemic fishes. *Knowledge & Management of Aquatic Ecosystems*, 417(36):01–07.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, 4. edition. ISBN 0-387-95457-0.

Verhulst, P.-F. (1838). Notice sur la loi que la population suit dans son accroissement. correspondance mathématique et physique publiée par a. *Quetelet*, 10:113–121.

Vilizzi, L., Stakenas, S., and Copp, G. H. (2012). Use of constrained additive and quadratic ordination in fish habitat studies: an application to introduced pumpkinseed Lepomis gibbosus and native brown trout Salmo truttain an English stream. *Fundamental and Applied Limnology*, 180(1):69–75.

Wang, Y., Naumann, U., Eddelbuettel, D., and Warton, D. (2018). *mvabund: Statistical Methods for Analysing Multivariate Abundance Data*. R package version 3.13.1.

Wang, Y. A., Naumann, U., Wright, S. T., and Warton, D. I. (2012). Mvabund- an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution*, 3(3):471–474.

Warton, D. I. (2008a). Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *Journal of the American Statistical Association*, 103(481):340–349.

Warton, D. I. (2008b). Raw data graphing: An informative but under-utilized tool for the analysis of multivariate abundances. *Australian Ecology*, 33(3):290–300.

Warton, D. I. (2011). Regularized Sandwich Estimators for Analysis of High-Dimensional Data Using Generalized Estimating Equations. *Biometrics*, 67(1):116–123.

Warton, D. I., Blanchet, F. G., Hara, R. B. O., Ovaskainen, O., Taskinen, S., Walker, S. C., and Hui, F. K. (2015a). So Many Variables: Joint Modeling in Community Ecology. *Trends in Ecology & Evolution*, 30(12):766–779.

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

Warton, D. I. and Hui, F. K. (2011). The arcsine is asinine: the analysis of proportions in ecology. *Ecology*, 92(1):3–10.

Warton, D. I. and Popovic, G. (2018). Model-based multivariate analysis. Presentation.

Warton, D. I., Thibaut, L., and Wang, Y. A. (2017). The PIT-trap - A model-free bootstrap procedure for inference about regression models with discrete, multivariate responses. *PLoS ONE*, 12(7):e0181790.

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

Westfall, P. and Young, S. (1993). *Resampling-Based Multiple Testing*. John Wiley & Sons, New York.

Yamamura, K. (1999). Transformation using (x+ 0.5) to stabilize the variance of populations. *Researches on Population Ecology*, 41(3):229–234.

Yee, T. W. (2004). A New Technique for Maximum-Likelihood Canonical Gaussian Ordination. *Ecological Monographs*, 74(4):685–701.

Yee, T. W. (2006). Constrained additive ordination. *Ecology*, 87(1):203–213.

Yee, T. W. (2015). *Vector generalized linear and additive models: with an implementation in R*. Springer, New York.

Yee, T. W. (2018). *VGAM: Vector Generalized Linear and Additive Models*. R package version 1.0-5.

Yee, T. W. and Hastie, T. J. (2003). Reduced-rank vector generalized linear models. *Statistical modelling*, 3(1):15–41.

Yee, T. W. and Wild, C. (1996). Vector generalized additive models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 481–493.

Zhang, Y. and Thas, O. (2012). Constrained ordination analysis in the presence of zero inflation. *Statistical Modelling*, 12(6):463–485.

Zuur, A. F. (1999). *Dimension reduction techniques in community ecology with applications to spatio-temporal marine ecological data*. PhD thesis, University of Aberdeen.

5 Supplementary Materials

5.1 Response Symmetry in GLMs

In the univariate significance tests of GLM$_{mv}$s, species with optima at the middle of an uni- or bimodal gradient were assigned high p-values. Here, I will show that this likely occurred due to the symmetry of the response shape when considered over the whole sampling range. Figure 7 shows three unimodal responses that only differ in the position of their optimum. One (orange) is at 50 which is exactly the middle of the sampled gradient, as is indicated by the dashed line, one (blue) is at 49 and the last (green) at 25.

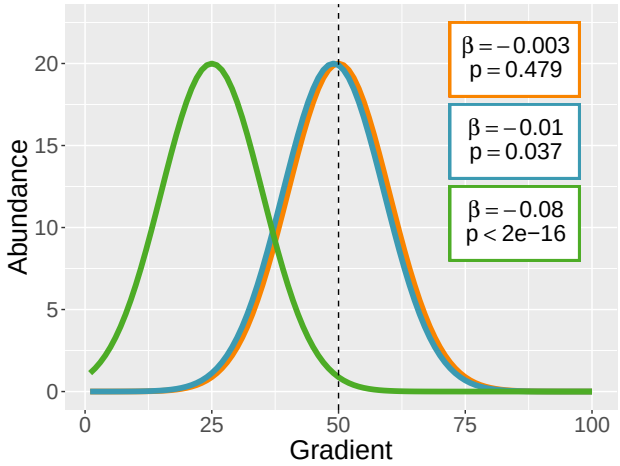


Figure 7: Unimodal response shapes with different optima. The dashed line indicates the middle of the sampled range. Each color represents a different species. The response shapes only differ in the location of their optimum, both tolerance and maximal abundance are equal. β refers to the regression coefficient for the gradient in a GLM which models abundance as a function of the gradient and p to the corresponding p-value. The color of the boxes indicates which curve the values belong to. The regression coefficient increase and the p-values decrease as the distance of the optimum from the middle of the gradient increases.

The p-values and regression coefficients come from a GLM with negative binomial residual distribution, conducted with the MASS R-package (Venables and Ripley, 2002). The small difference in optima between the orange and the blue species entails a big disparity in regression coefficients. The regression coefficient of the blue species is approximately three times higher and the p-value is less than a tenth of the orange ones. The green species shows that the regression coefficient further increases and the p-value further decreases as the optimum is farther removed from the middle. These results were comparable but not identical to those obtained when using optima higher than 50 (not shown here).

Figure 8 shows a similar setup for bimodal response shapes. In figure 8a the response is symmetrical. The middle of the gradient lies exactly between the two optima. A GLM with negative binomial residual distribution returns a low regression coefficient (- 9e-04) and an appropriately high p-value

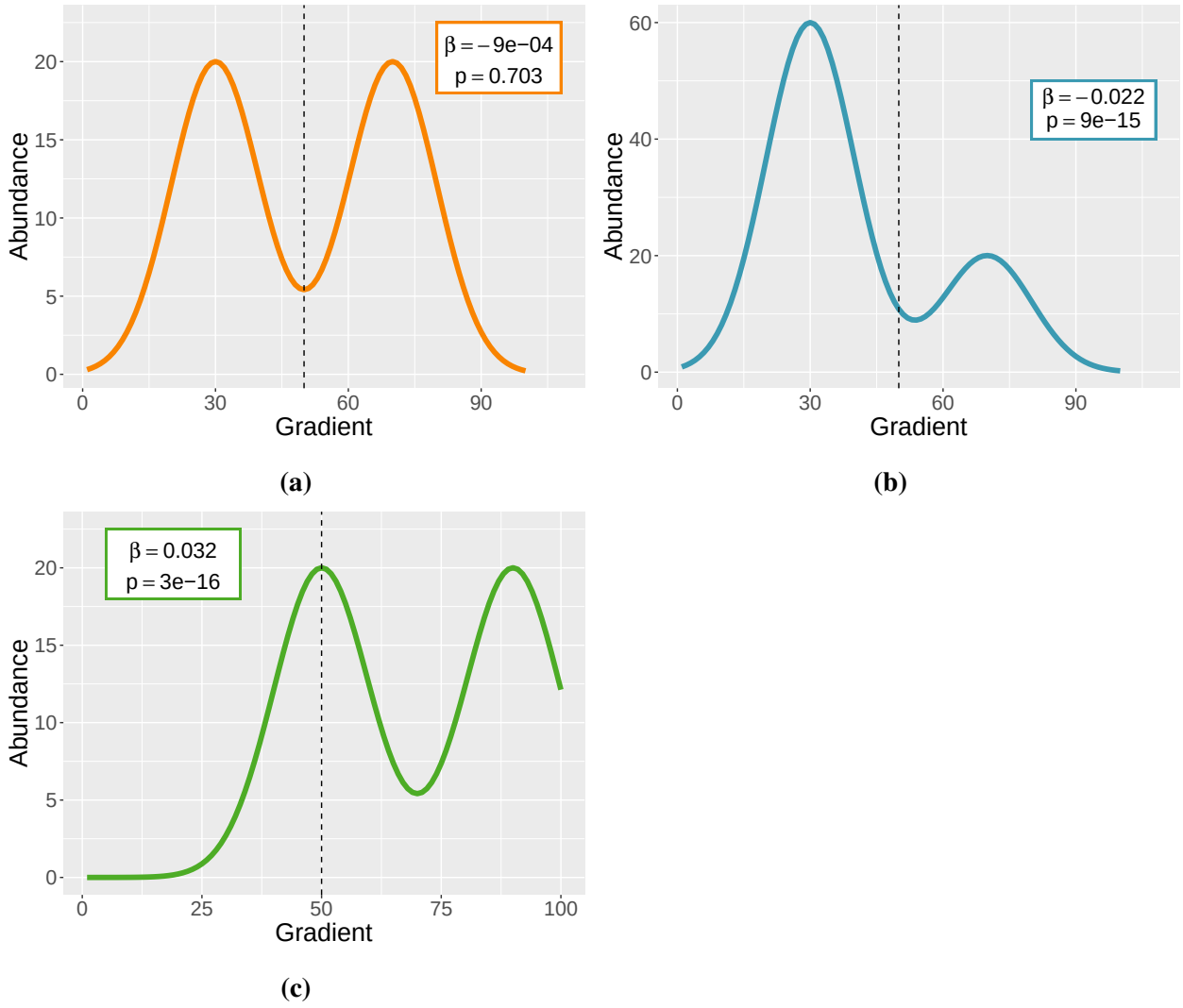


Figure 8: Bimodal response shapes. In the boxes β is the regression coefficient for the gradient in a GLM which models Abundance as a function of the gradient. p is the associated p-value. The dashed line indicates the middle of the sampled range. (**a**) shows a symmetrical response, β is low and the p-value is high. (**b**) and (**c**) are asymmetrical modifications of (**a**). In (**b**) the maximum of the first optimum is higher than the second. This results in a higher β and lower p-value. (**c**) shows the same response shape as (**a**) but shifted to the right. This modification also results in a higher β and lower p-value.

(0.703). For figure 8b, the maximal abundance at the first optimum was increased to three times that of the second. The middle of the response shape still coincides with that of the gradient. The regression coefficient (-0.22) is increased (in absolute terms) relative to the first model and the p-value (9e-15) is decreased. A similar pattern can be observed when the response shape of figure 8a is shifted to the right, as is done in figure 8c. Again, the regression coefficient (0.032) is increased and the p-value (3e-16) decreased.

These examples show that the symmetry of the response shape over the whole sampling range leads to small regression coefficients and the corresponding high p-values. They also demonstrated that

already small deviations from this symmetry (1/100 of the gradient length as the blue species in Figure 7) substantially increases the coefficient and decreases the p-value.

5.2 Further Details on Simulations

The Table 3 shows the model parameters used in the simulation.

Table 3: Model parameters used in simulations. An x indicates that the parameter is not relevant to the gradient types used. c is the maximal abundance, t the tolerance, u the location of the optimum, x_0 the midway points of the sigmoid, β the linear response parameter and k the steepness of the logistic curve. Values in braces are optima pairs for bimodal gradients.

	c	t	u/ x_0	β	k
Uni-Uni	100	7.5	20, 50, 80	x	x
Uni-Li	100	7.5	0, 25, 50, 75, 100	1, 1.2, 1.4, 1.6, 1.8	x
Uni-Lo	100	7.5	20, 50, 80	x	0.1
Uni-Bi	100	5	20, 50, 80/ {10, 30}, {40, 60}, {70, 90}	x	x
Li-Li	x	x	x	80, 100, 120, 0.8, 1.1, 1.2	x
Li-Lo	100	x	15, 30, 45, 60, 75	0.1	0.05
Li-Bi	100	6	{5, 25}, {25, 45}, {35, 55}, {55, 75}, {75, 95}	0.1	x
Lo-Lo	100	x	20, 50, 80	x	0.1
Lo-Bi	100	6	20, 50, 80/ {5, 25}, {35, 55}, {75, 95}	x	0.1
Bi-Bi	100	6	{5, 25}, {35, 55}, {75, 95}	x	x

The optimum parameter u is the only instance of a parameter that is relevant to both gradients and differing between them. The different values are separated by forward slashes. Bimodal gradients require two optima per species. The used combinations are shown in braces.
In *Li-Li*, the first three species reacted to *env1* the next three to *env2* and the last three to both. For this reason, the β parameters of the last three species are much lower than the first three. This also causes the high p-values for *env1* and *env2* in *Li-Li* (see Table 6).

5.3 Ordination Diagrams

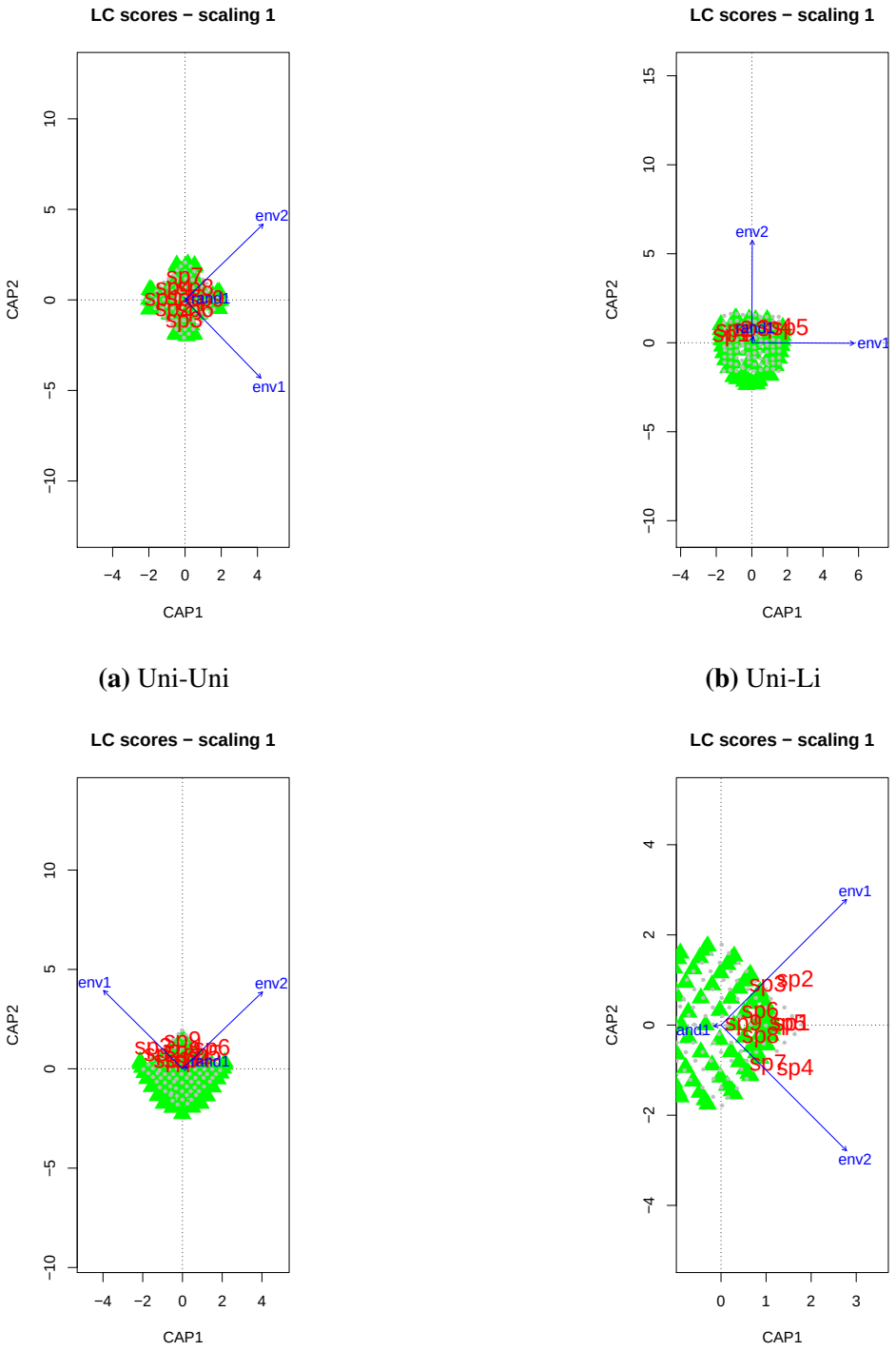
As the result of dimension reduction, db-RDA, CCA, and CAO/CQO produce ordination diagrams. In the following section, a selection of diagrams will be presented for all four methods.
This selection will include the models: *Uni-Uni*, *Uni-Li*, *Li-Li* and *Lo-Lo*. The first diagrams are from the db-RDA analyses (Figure 9). All subplots include explanatory variables (blue arrows), species scores (red letters), site scores as linear combinations of explanatory variables (grey points, LC-scores) and sites scores as weighted averages of species scores (green triangles, WA-scores). These data are plotted in scaling 1 (distance-triplot) and can be interpreted as follows: i.) the angles between explanatory variables and the arrow that could be drawn from the centroid to each

species score represent their correlation, ii.) distances between objects approximate their Euclidean distance.

The species scores are positioned correctly in Figure 9 a-c but in d the original pattern is not recognizable anymore. LC-scores always form squares and the distance between the points is approximately constant. WA-scores are more prone to deformation. In 9 a WA-scores are concentrated around species scores, in b they are fan-shaped, in c they look good and in d they appear to form a circle instead of a square. In all models explanatory variables either load equally strong on both axes or each load on a separate axis.

Figure 10 shows the triplots of the CCA analyses. The same symbology as in Figure 9 was used. Again, scaling 1 is utilized. In CCA this can be interpreted following these rules: i.) An orthogonal projection from an object (site or species score) on an explanatory variable approximates that object's position along that variable, ii.) distances among object approximate their chi-squared distance.

In Figure 10 a,c and d the species scores are placed correctly relative to the explanatory variables. In *Uni-Li* they are only ordered along the unimodal gradient and the second axis is determined by noise variables. The LC-scores form are square and are equidistant in *Uni-Uni* and *Lo-Lo*. In *Uni-Li* they appear to be randomly distributed and in *Li-Li* they are fan-shaped with the distance between points increasing towards lower values of the explanatory variables. WA-scores are concentrated around species scores in *Uni-Uni*, fan-shaped in *Li-Li* and *Lo-Lo* and scattered, though mostly absent in *Uni-Li*. In contrast to the db-RDA triplots, the species scores never lie outside the site scores.



(a) Uni-Uni

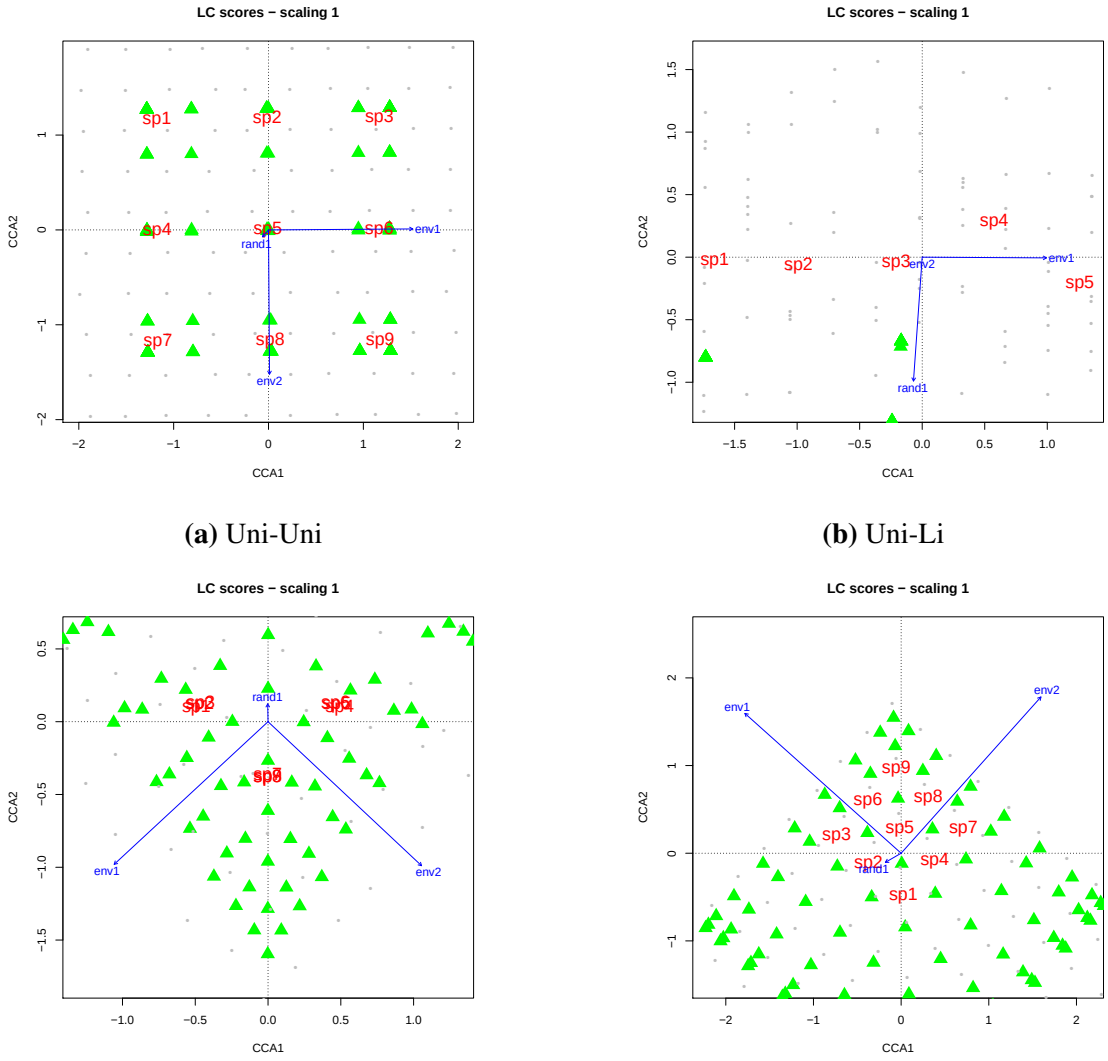
(b) Uni-Li

(c) Li-Li

(d) Lo-Lo

Figure 9: Triplots of the db-RDA analyses of (**a**) Uni-Uni, (**b**) Uni-Li, (**c**) Li-Li and (**d**) Lo-Lo. For every response combination the class 1 model was used.

Figure 11 shows the latent variables plots of the CAO analyses. The x-axis shows the value of the latent variable and the y-axis the estimated abundance. Each curve represents one species and is identifiable via the labels next to it. The plot for *Uni-Uni* in Figure 11a depicts the species as



(a) Uni-Uni

(b) Uni-Li

(c) Li-Li

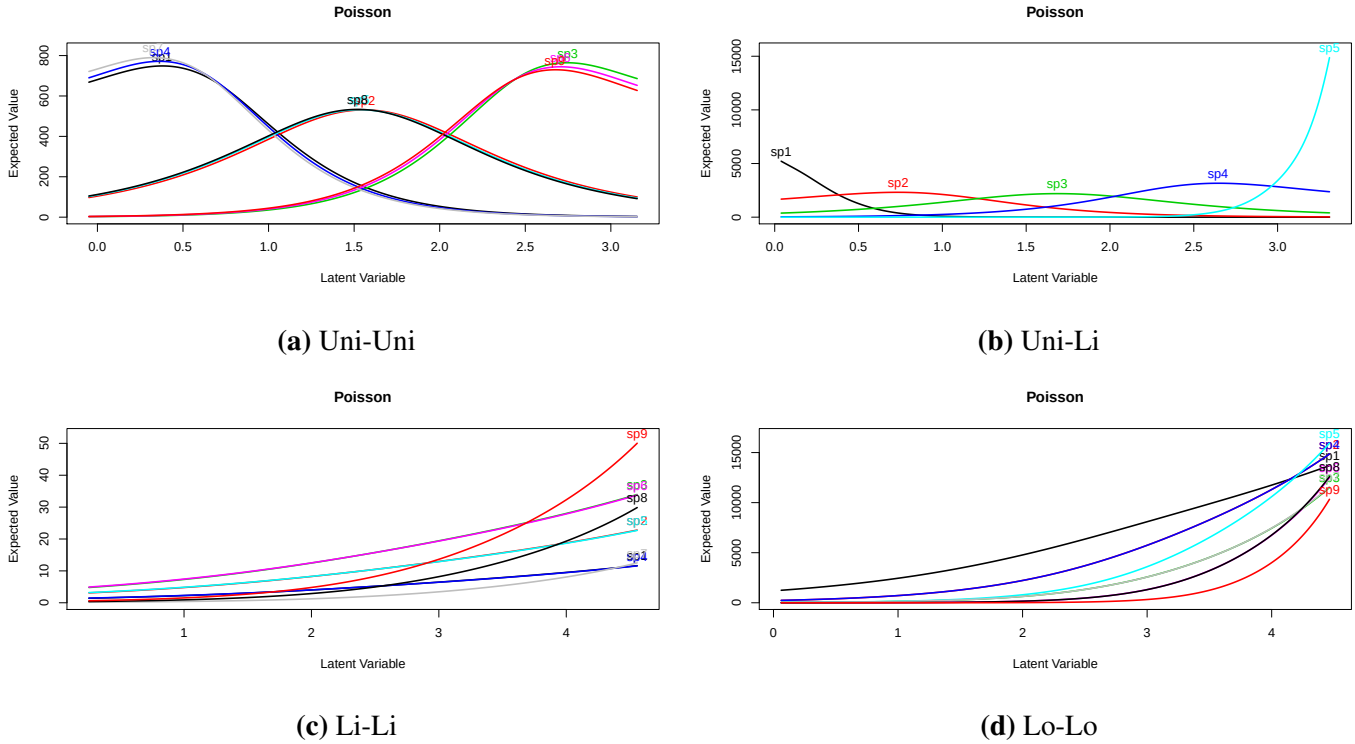
(d) Lo-Lo

Figure 10: Triplots of the CCA analyses of (**a**) Uni-Uni, (**b**) Uni-Li, (**c**) Li-Li and (**d**) Lo-Lo. For every response combination the class 1 model was used.

belonging to one of three groups based on their optimum. Based on the species identities (1, 4, 7 in group 1; 2, 5, 8 in group 2; 3, 6, 9 in group 3) it can be inferred that the latetent variable is only structured along *env1*. The species belonging to group 2 have lower maximal abundances than those in group 1 and 3, but all species lie below their actual maxima, which is 1000 for every species. In *Uni-Li* species are only ordered along the unimodal gradient *env1*. The abundances of species 1 and 5, which have their optima outside of the sampled range, are markedly higher than those of the other species and also higher than the model maxima. These *edge effects* in *Uni-Uni* and *Uni-Li* are common in CAO and CQO. Estimation of species with parts of their response outside the latent variable space failed more often and were the least accurate. Yee (2015) advises to drop the corresponding species form the analysis.

In *Li-Li* the smoothing induces a slight curve in the response shape, especially in the species 7 to 9.

The equality of species 1 and 4, 2 and 5 as well as 3 and 6 implies that both gradients weight equally strong on the latent variable. In Figure 11d none of the species reach their maximum despite all species, except 7 to 9, reaching theirs in the model. Again, the equivalence of the curves for species 2 and 4, 3 and 7 as well as 6 and 8 suggests that both explanatory variables weigh identically on the latent variable.



(a) Uni-Uni

(b) Uni-Li

(c) Li-Li

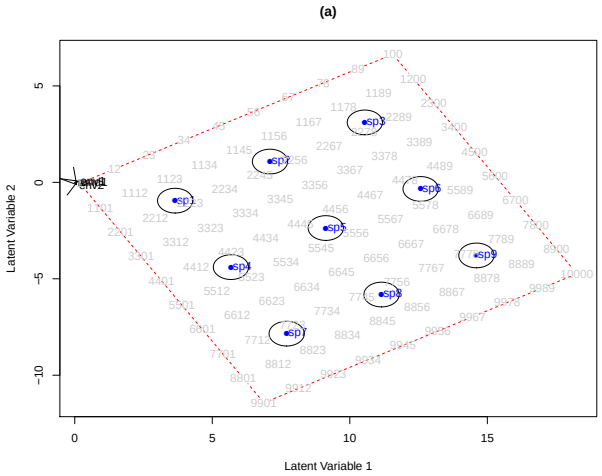
(d) Lo-Lo

Figure 11: Latent variable plots of the CAO analyses of class 1 (**a**) Uni-Uni, (**b**) Uni-Li, (**c**) Li-Li and (**d**) Lo-Lo.

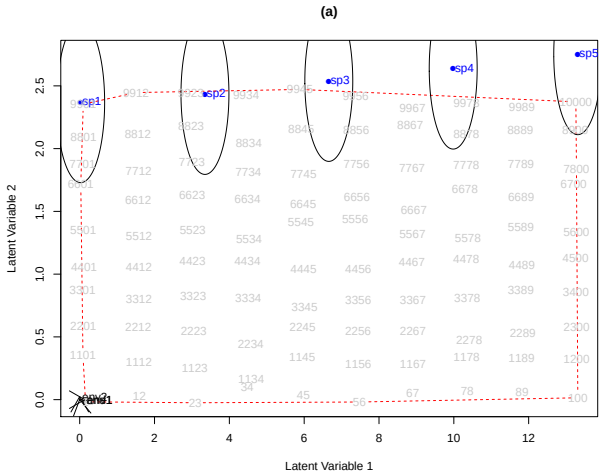
The latent variable plots of the CQO analyses are presented in Figure 12. The plots show the equal tolerance models. The red line bounds the site scores and represents the latent variables space which corresponds to the measured range (the convex hull). Grey numbers are the site scores and blue points indicate species optima. The circles around optima are areas where the abundance is above 95% of its maximum. Explanatory variables should be represented by black arrows, but they are too short to be interpretable here.

Uni-Uni is represented flawlessly in Figure 12a. For *Uni-Li* the species are ordered along *env1* and optima slightly move up on the latent variable 2 from species 1 to 5. Since the measured optimum of every species is at the same value of *env2* (which is represented by the latent variable 2) this increase reflects an affect of abundance on the estimation of the optimum. The effect is only weak though. In *Li-Li* all species have their optima outside the convex hull. According to Yee (2015) such species should be removed from the model, as there is a lot of uncertainty associated with them. However, here the estimates are good. The arrangement of species and sites in the latent variables space is close to the model. Lastly, in *Lo-Lo* most species are again outside the convex hull. Both

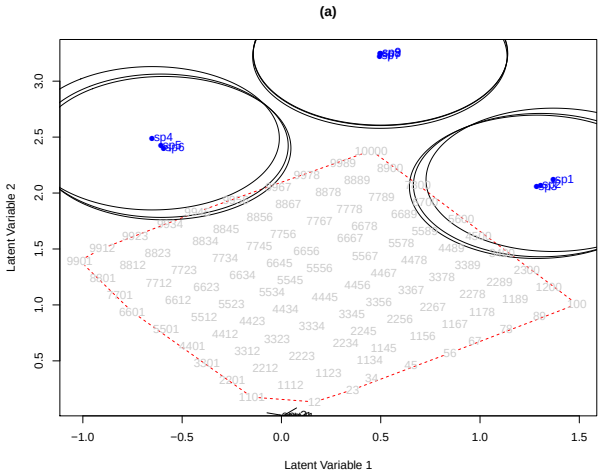
the array of sites and species is good, it is just shifted relative to each other. The position of optima on logistic gradients seem to be categorically biased towards higher latent variable values.



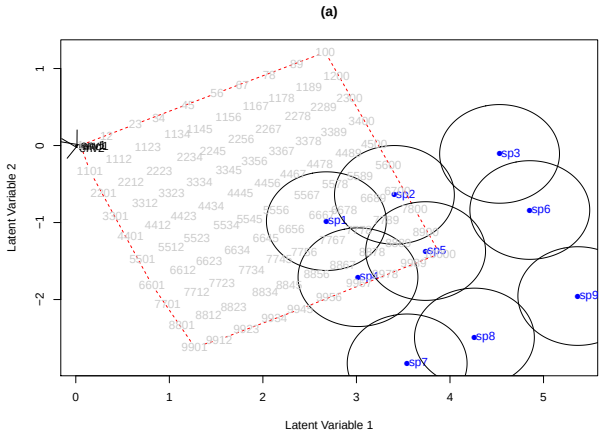
(a) Uni-Uni



(b) Uni-Li



(c) Li-Li



(d) Lo-Lo

Figure 12: Latent variable plots of the rank-2 CQO analyses of (**a**) Uni-Uni, (**b**) Uni-Li, (**c**) Li-Li and (**d**) Lo-Lo. For every response combination the class 1 model was used.

5.4 Further Result Statistics

This section contains all the p-values, constrained coefficients, inertias and differences in maxima estimates at the level of response types or classes.
Tables 4 to 7 show the p-values of GLM$_{mv}$s. The Tables 4 and 5 contain the multivariate p-values and the Tables 6 and 7 hold the univariate p-values.

Table 4: Multivariate p-values of GLM$_{mv}$s at the level of classes. Class four is missing as no models of this class converged. For explanation of classes see Table 1.

	env1		env2		Noise	
	μ	σ	μ	σ	μ	σ
Class 1	0.002	6e-4	0.003	0.002	0.407	0.235
Class 2	0.002	6e-4	0.003	0.002	0.565	0.242
Class 3	0.002	6e-4	0.002	8e-4	0.615	0.286

Table 5: Multivariate p-values of GLM$_{mv}$s at the level of response types.

	env1		env2		Noise	
	μ	σ	μ	σ	μ	σ
Uni-Uni	0.002	0	0.002	0	0.666	0.205
Uni-Li	0.002	0	0.002	9e-4	0.589	0.242
Uni-Lo	0.002	0	0.002	7e-4	0.591	0.241
Uni-Bi	0.002	0	0.006	0.002	0.677	0.164
Li-Li	0.002	0	0.002	0	0.665	0.312
Li-Lo	0.003	0.001	0.003	0.001	0.482	0.311
Li-Bi	0.002	0	0.002	0	0.405	0.294
Lo-Lo	0.002	0	0.002	0	0.521	0.317
Lo-Bi	0.002	0	0.002	0	0.597	0.269
Bi-Bi	0.002	0	0.004	0.003	0.612	0.191

Table 6: Univariate p-values of GLM$_{mv}$s at the level of classes. Class four is missing as no models of this class converged.

	env1		env2		Noise	
	μ	σ	μ	σ	μ	σ
Class 1	0.179	0.354	0.132	0.313	0.754	0.287
Class 2	0.171	0.345	0.122	0.302	0.790	0.228
Class 3	0.155	0.329	0.118	0.317	0.821	0.240

Table 7: Univariate p-values of GLM$_{mv}$s at the level of response types.

	env1		env2		Noise	
	μ	σ	μ	σ	μ	σ
Uni-Uni	0.332	0.470	0.333	0.472	0.901	0.135
Uni-Li	0.139	0.282	0.010	0.005	0.805	0.213
Uni-Lo	0.301	0.428	0.034	0.059	0.827	0.187
Uni-Bi	0.332	0.470	0.339	0.470	0.894	0.129
Li-Li	0.280	0.396	0.324	0.459	0.842	0.230
Li-Lo	0.003	0.003	0.003	0.002	0.574	0.312
Li-Bi	0.003	0.002	0.003	0.002	0.585	0.312
Lo-Lo	0.003	0.002	0.002	0	0.718	0.308
Lo-Bi	0.007	0.009	0.004	0.004	0.833	0.199
Bi-Bi	0.141	0.215	0.053	0.092	0.853	0.146

Tables 8 to 11 show the p-values of dbRDAs. The Tables 8 and 9 contain the p-values for terms and the Tables 10 and 11 hold the p-values for axes.

Table 8: p-values of terms in db-RDAs at the level of response types.

	env1		env2		Noise	
	μ	σ	μ	σ	μ	σ
Uni-Uni	0.001	0	0.001	0	0.479	0.306
Uni-Li	0.001	0	0.067	0.119	0.475	0.264
Uni-Lo	0.001	0	0.072	0.130	0.508	0.254
Uni-Bi	0.001	3e-4	0.001	0	0.523	0.298
Li-Li	0.001	0	0.001	0	0.421	0.160
Li-Lo	0.001	0	0.001	9e-4	0.446	0.235
Li-Bi	0.002	0.002	0.001	0	0.459	0.243
Lo-Lo	0.001	0	0.001	0	0.383	0.206
Lo-Bi	0.002	0.002	0.001	0	0.487	0.237
Bi-Bi	0.001	0	0.001	0	0.529	0.313

Table 9: *p*-values of terms in db-RDAs at the level of classes. For explanation of classes see Table 1.

	env1		env2		Noise	
	μ	σ	μ	σ	μ	σ
Class 1	0.001	0	0.001	0	0.466	0.220
Class 2	0.001	0	0.001	0	0.478	0.245
Class 3	0.001	0	0.001	0	0.431	0.238
Class 4	0.002	0.002	0.056	0.112	0.506	0.289

Table 10: *p*-values of axes in db-RDAs at the level of response types.

	CAP1		CAP2		CAP3	
	μ	σ	μ	σ	μ	σ
Uni-Uni	0.004	0.005	0.012	0.023	0.575	0.286
Uni-Li	0.001	3e-4	0.185	0.358	0.511	0.276
Uni-Lo	0.001	0	0.180	0.345	0.673	0.274
Uni-Bi	0.005	0.008	0.079	0.159	0.570	0.288
Li-Li	0.001	0	0.001	0	0.670	0.204
Li-Lo	0.001	0	0.063	0.107	0.866	0.127
Li-Bi	0.001	0	0.013	0.022	0.452	0.248
Lo-Lo	0.001	0	0.044	0.079	0.557	0.331
Lo-Bi	0.001	0	0.035	0.063	0.612	0.293
Bi-Bi	0.001	0	0.001	4e-4	0.556	0.291

Table 11: *p*-values of axes in db-RDAs at the level of classes.

	CAP1		CAP2		CAP3	
	μ	σ	μ	σ	μ	σ
Class 1	0.001	0	0.001	4e-4	0.572	0.280
Class 2	0.001	0	0.003	0.006	0.571	0.282
Class 3	0.001	0	0.001	0	0.502	0.297
Class 4	0.004	0.006	0.240	0.288	0.773	0.174

Tables 12 to 17 show the *p*-values of CCAs. The Tables 12 and 13 contain the *p*-values for terms, the Tables 14 and 15 hold the *p*-values for axes and the Tables 16 and 17 contain the inertias.

Table 12: *p*-values of terms in CCA at the level of classes. For explanation of classes see Table 1.

	env1		env2		Noise	
	μ	σ	μ	σ	μ	σ
Class 1	0.201	0.405	0.101	0.305	0.389	0.264
Class 2	0.201	0.406	0.101	0.305	0.336	0.293
Class 3	0.201	0.406	0.101	0.305	0.404	0.320
Class 4	0.202	0.406	0.102	0.304	0.312	0.315

Table 13: *p*-values of terms in CCA at the level of response types.

	env1		env2		Noise	
	μ	σ	μ	σ	μ	σ
Uni-Uni	0.001	0	0.001	0	0.155	0.251
Uni-Li	0.001	0	1	0	0.302	0.274
Uni-Lo	0.001	0	0.004	0.008	0.207	0.247
Uni-Bi	0.001	0	0.001	0	0.118	0.193
Li-Li	0.001	0	0.001	0	0.544	0.236
Li-Lo	0.998	0.002	0.001	0	0.388	0.263
Li-Bi	1	0.000	0.001	0	0.445	0.323
Lo-Lo	0.001	0	0.001	0	0.414	0.290
Lo-Bi	0.004	0.006	0.001	0	0.452	0.318
Bi-Bi	0.001	0	0.001	0	0.507	0.336

Table 14: *p*-values of axes in CCA at the level of classes.

	CCA1		CCA2		CCA3+	
	μ	σ	μ	σ	μ	σ
Class 1	0.001	0	0.223	0.364	0.528	0.370
Class 2	0.001	0	0.086	0.227	0.778	0.360
Class 3	0.001	0	0.155	0.299	0.799	0.336
Class 4	0.001	0	0.118	0.188	0.678	0.403

Table 15: p-values of axes in CCA at the level of response types.

	env1		env2		Noise	
	μ	σ	μ	σ	μ	σ
Uni-Uni	0.001	0	0.001	0	0.364	0.417
Uni-Li	0.001	0	0.341	0.239	0.850	0.165
Uni-Lo	0.001	0	0.009	0.016	0.721	0.370
Uni-Bi	0.001	0	0.001	0	0.407	0.407
Li-Li	0.001	0	0.001	5e-4	0.999	0.003
Li-Lo	0.001	0	0.633	0.398	0.999	8e-4
Li-Bi	0.001	0	0.403	0.304	0.965	0.060
Lo-Lo	0.001	0	0.001	0	0.799	0.329
Lo-Bi	0.001	0	0.066	0.138	0.840	0.298
Bi-Bi	0.001	0	0.001	6e-4	0.807	0.278

Table 16: Inertia of CCAs at the level of classes.

	Total		Constrained		Unconstrained	
	μ	σ	μ	σ	μ	σ
Class 1	2.97	2.57	1.01	0.67	1.96	1.92
Class 2	2.97	2.57	1.16	0.81	1.81	1.78
Class 3	2.97	2.59	1.04	0.72	1.92	1.90
Class 4	3.21	2.87	1.99	1.85	1.22	1.20

Table 17: Total, Constrained and Unconstrained Inertia of CCA at the level of response types.

	Total		Constrained		Unconstrained	
	μ	σ	μ	σ	μ	σ
Uni-Uni	7.30	0.34	3.04	1.63	4.26	1.29
Uni-Li	3.51	0.26	1.29	0.43	2.22	0.18
Uni-Lo	2.36	0.07	1.31	0.32	1.06	0.25
Uni-Bi	6.75	0.51	2.75	1.24	4.00	0.75
Li-Li	0.28	0.04	0.23	0.03	0.05	0.01
Li-Lo	0.05	0.00	0.05	0.00	0.00	0.00
Li-Bi	1.62	0.08	0.90	0.10	0.73	0.18
Lo-Lo	0.36	0.00	0.30	0.01	0.07	0.01
Lo-Bi	2.08	0.01	1.12	0.13	0.97	0.14
Bi-Bi	5.96	0.06	2.02	0.37	3.94	0.39

For CAOs and CQOs the difference in maxima is also considered. In CQOs the tolerance setting is used as an additional level to summarize the data.
Tables 18 to 21 refer to CAOs. The Tables 18 and 19 contain the constrained coefficients and the Tables 20 and 21 hold the difference between estimated and actual model maxima, expressed in percent of the respective model maximum.

Table 18: Constrained Coefficients of CAO at the level of response types.

	env1		env2		Noise	
	μ	σ	μ	σ	μ	σ
Uni-Uni	0.007	0.013	0.023	0.014	0.001	0.003
Uni-Li	0.032	0.001	0.002	1e-4	2e-4	2e-4
Uni-Lo	0.031	0.002	2e-4	2e-4	0.001	4e-4
Uni-Bi	0.032	0.001	4e-4	3e-4	0.001	4e-4
Li-Li	0.022	0.002	0.022	0.003	0.001	0.001
Li-Lo	0.024	0.001	0.019	0.002	0.001	0.001
Li-Bi	0.001	2e-4	0.031	0.002	4e-4	3e-4
Lo-Lo	0.022	0.002	0.022	0.001	0.001	0.001
Lo-Bi	0.002	0.003	0.031	0.003	0.001	0.001
Bi-Bi	0.017	0.008	0.025	0.006	0.001	0.001

Table 19: Constrained Coefficients of CAO at the level of classes. For explanation of classes see Table 1.

	env1		env2		Noise	
	μ	σ	μ	σ	μ	σ
Class 1	0.020	0.013	0.017	0.013	0.0004	4e-4
Class 2	0.018	0.013	0.019	0.013	0.0006	5e-4
Class 3	0.020	0.013	0.018	0.013	0.0004	3e-4
Class 4	0.015	0.011	0.018	0.009	0.0015	0.002

Table 20: Difference between model maxima and estimated maxima of CAO expressed in % of the respective model maximum, at the level of response types.

	Δ Maximum	
	μ	σ
Uni-Uni	-92.65	2.29
Uni-Li	-81.51	3.25
Uni-Lo	-79.59	9.30
Uni-Bi	-93.49	1.27
Li-Bi	-74.79	3.13
Lo-Bi	-75.99	10.79
Bi-Bi	-86.02	3.68

Table 21: Difference between model maxima and estimated maxima of CAO expressed in % of the respective model maximum, at the level of classes.

	Δ Maximum	
	μ	σ
Class 1	-86.28	8.77
Class 2	-86.10	8.75
Class 3	-85.56	9.06
Class 4	-78.95	9.63

Tables 22 to 27 refer to CQOs. The Tables 22 to 24 contain the constrained coefficients and the Tables 25 to 27 hold the difference between estimated and actual model maxima, expressed in percent of the respective model maximum.

Table 22: Constrained Coefficients of CQO at the level of response types

	env1		env2		Noise	
	μ	σ	μ	σ	μ	σ
Uni-Uni	0.170	0.031	0.168	0.032	0.004	0.006
Uni-Li	0.091	0.050	0.031	0.009	0.004	0.004
Uni-Lo	0.128	0.018	0.028	0.006	0.002	0.002
Uni-Bi	0.182	0.060	0.082	0.018	0.003	0.002
Li-Li	0.028	0.007	0.028	0.006	0.002	0.002
Li-Lo	0.028	0.003	0.022	0.007	0.002	0.003
Li-Bi	0.031	0.009	0.080	0.055	0.003	0.003
Lo-Lo	0.036	0.004	0.036	0.004	0.001	0.001
Lo-Bi	0.031	0.006	0.077	0.020	0.002	0.001
Bi-Bi	0.106	0.017	0.106	0.017	0.002	0.002

Table 23: Constrained Coefficients of CQO at the level of classes.

	env1		env2		Noise	
	μ	σ	μ	σ	μ	σ
Class 1	0.085	0.065	0.067	0.045	0.001	0.002
Class 2	0.080	0.062	0.062	0.047	0.002	0.002
Class 3	0.082	0.064	0.064	0.046	0.001	0.002
Class 4	0.065	0.055	0.061	0.061	0.004	0.005

Table 24: Constrained Coefficients of CQO at the level of tolerance settings.

	env1		env2		Noise	
	μ	σ	μ	σ	μ	σ
equal	0.079	0.062	0.061	0.049	0.002	0.003
identical	0.084	0.062	0.065	0.044	0.003	0.003
unequal	0.074	0.063	0.064	0.054	0.002	0.003

Table 25: Difference between model maxima and estimated maxima of CQO expressed in % of the respective model maximum, at the level of response type.

	Δ Maximum	
	μ	σ
Uni-Uni	2310,94	23748,73
Uni-Li	1,64e+08	1,55e+09
Uni-Lo	3,00e+30	4e+031
Uni-Bi	2e+096	2e+097
Li-Li	311,58	1313,77
Li-Lo	5,76	38,03
Li-Bi	2e+053	1e+054
Lo-Lo	3,23e+14	5,28e+15
Lo-Bi	2e+029	2e+030
Bi-Bi	1867,95	16442,15

Table 26: Difference between model maxima and estimated maxima of CQO expressed in % of the respective model maximum, at the level of classes.

	Δ Maximum	
	μ	σ
Class 1	8e+054	2e+056
Class 2	7e+023	1e+025
Class 3	6e+095	1e+097
Class 4	5e+052	7e+053

Table 27: Difference between model maxima and estimated maxima of CQO expressed in % of the respective model maximum, at the level of tolerance settings.

	Δ Maximum	
	μ	σ
equal	5e+095	9e+096
identity	2e+026	4e+027
unequal	3e+052	5e+053