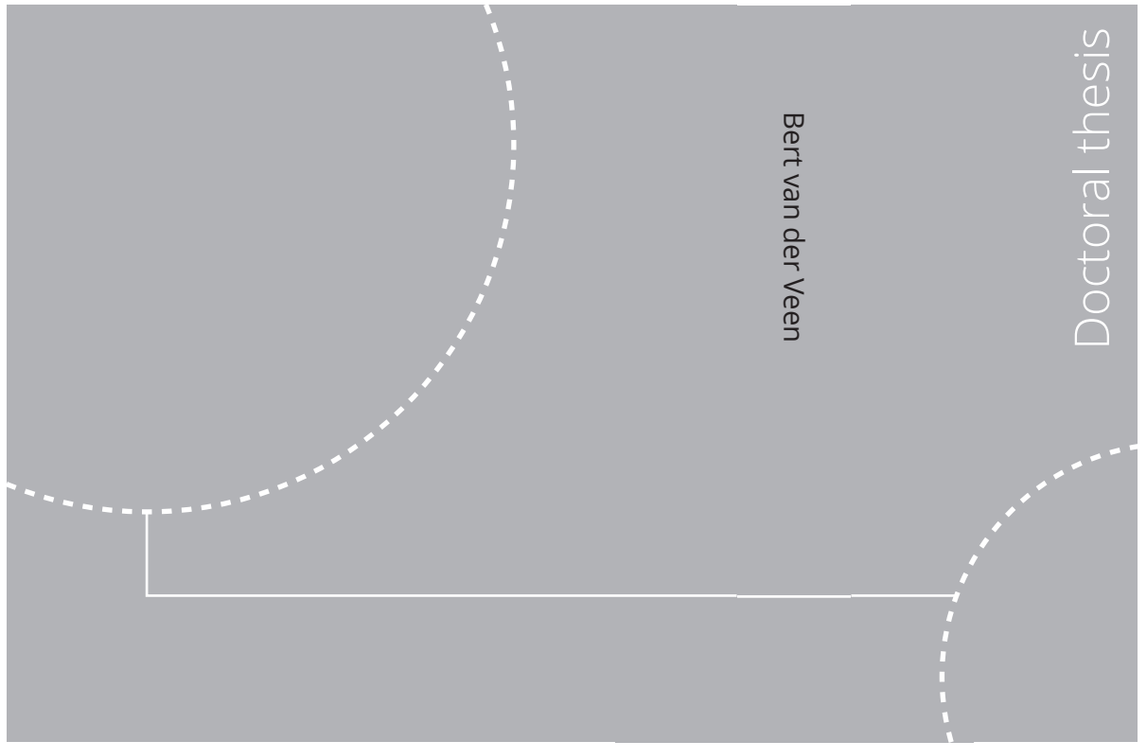
ISBN 978-82-326-6478-8 (printed ver.)
ISBN 978-82-326-5610-3 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (electronic ver.)



Doctoral theses at NTNU, 2022:50

Bert van der Veen

Statistical advances in multispecies models for community ecology

NTNU
Norwegian University of
Science and Technology
Thesis for the degree of
Philosophiae Doctor
Faculty of Information Technology
and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Doctoral theses at NTNU, 2022:50

NTNU

NTNU
Norwegian University of
Science and Technology

Bert van der Veen

Statistical advances in multispecies models for community ecology

Thesis for the degree of Philosophiae Doctor

Trondheim, February 2022

Norwegian University of Science and Technology
Faculty of Information Technology
and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of Science and Technology

Thesis for the degree of Philosophiae Doctor

Faculty of Information Technology
and Electrical Engineering
Department of Mathematical Sciences

ISBN 978-82-326-6478-8 (printed ver.)
ISBN 978-82-326-5610-3 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (electronic ver.)

Doctoral theses at NTNU, 2022:50

PRINTED IN NORWAY
NO - 1598

NORDIC SWAN ECOLABEL
Printed matter
2041 0731

Printed by Skipnes Kommunikasjon AS

This was supposed to be a PhD in data integration. Oh well.

Acknowledgements

The completion of this PhD has been made possible due to the significant contributions of Bob, Knut, and Francis. Bob intensively supported me, especially in the last year, and tolerated my uninhibited spam on Skype. He taught me how to develop ideas by providing me with the freedom to pursue topics that interest me. But, also that almost all models are really a type of Generalized Linear Model anyway. Together with Bob, Knut initiated my employment, and I am forever grateful for having had this opportunity to dive further into the rabbit hole that is statistical modelling. Especially in the beginning of my PhD, Knut made me feel welcome, which greatly motivated me. Although he terminated his employment with NIBIO at the beginning of my second year, he still managed to find the time for my supervision while employed at Artsdatabanken. Francis stepped up as additional co-supervisor when I needed a great deal of technical support. He patiently taught me the more mathematical aspects of statistics, which has made it possible for me to understand the nitty gritty details of GLLVMs. I am confident in saying that I could not have written this thesis without any of them.

If it had not been for the pandemic, I would have gone to Australia in order to collaborate with David Warton, and to meet Francis in person. Unfortunately that did not happen. Regardless, I look forward to potential future collaborations on model-based ordination or related subjects!

NIBIO financed my PhD education and provided me with a research group. During the covid-19 pandemic NIBIO supported me with additional funding, when it was clear that I was struggling. For this, I thank Anders Nielsen. I thank Erik Solbu for making this experience more enjoyable. He was always available for a chat and for advice, or for a board game. In general, my colleagues in NIBIO provided support by listening to my rants on statistics.

I want to thank Sam Perrin and Jenni Niku for their role in organising two workshops on GLLVMs for ecologists: one at the Festival of Ecology by the British Ecological Society, and one at the International Summer School in Applied Ecology in Evenstad. Method developments are useless if the audience they are intended for is not educated in their application. Furthermore, I want to thank Jenni, but also Sara Taskinen, for allowing me to take part in developing the `gllvm` `R`-package, and for providing a platform to disseminate the methods in this thesis.

And finally, I want to thank my girlfriend Maris for standing by my side. She was patient with me beyond measure, even when she started her own PhD education. Patience is not one of my virtues, and I truly hope that during the remainder of her PhD education I can manage to

be even half as patient as she was with me. Anyone reading this should imagine me ranting on about statistically minded subjects, or the abuse of statistical methods by ecologists, which has the tendency to frustrate me incredibly. Even though Maris often did not understand what I was talking of, she always waited until I was finished, after which I would ask her "does that make sense to you?" and she would simply answer "no", and smile.

Preface

In 2017 I arrived in snowy Stjørdal after a long drive from Evenstad. Knut was waiting to welcome me at the NIBIO research station when I got stuck in the snow with my car. A messy hour with Knut on a tractor later, and I was finally inside. The first year was an eventful year, since the research group was moved from Stjørdal to Trondheim.

Since the start of my PhD a lot has happened. On my second day, I started a course at NTNU in "advanced computer intensive statistical methods" taught by prof. Jo Eidsvik. That was my trial-by-fire introduction to statistics. I would say it went pretty well, though it was more than challenging at times. I do believe that my background in ecology, before doing a PhD in statistics, has given me an unique perspective.

In this thesis, my goal has been to provide suitable alternatives to classical ordination methods for ecologists. In the end, it turns out to be a collection of thoughts and developments that relate to both ordination and species distribution modelling.

This thesis is divided into three different sections, including a total of seven articles. The first section includes developments of methods for community ecologists. The second section focuses on two ecological applications of the GLLVM framework, the first using (joint) species distribution modelling, and the second using (model-based) ordination. The model developments in the first section, and the applications of the second section, can be seen in light of either ordination or species distribution modelling.

In general, the choice of JSDM versus ordination is an arbitrary one, which can be demonstrated quite straightforwardly. For example, in the first example, which includes the modelling of species associations in a freshwater community as a function of temperature, we visualize species associations in the form of residual correlations. In an article more focused on using ordination for community ecology, an ordination plot might have been used instead, which we can still do post-hoc! The figure below visualizes species associations as a function of temperature, but using ordination plots:

The second example of a plant-pollinator system in Trondheim chooses to model and visualize patterns in the data differently. Here, sites are modelled as a function of the environment rather than species, so that a different ordination plot of sites can be constructed at each point along the gradient. Alternatively, a single plot of species associations as represented by residual correlations can be drawn:

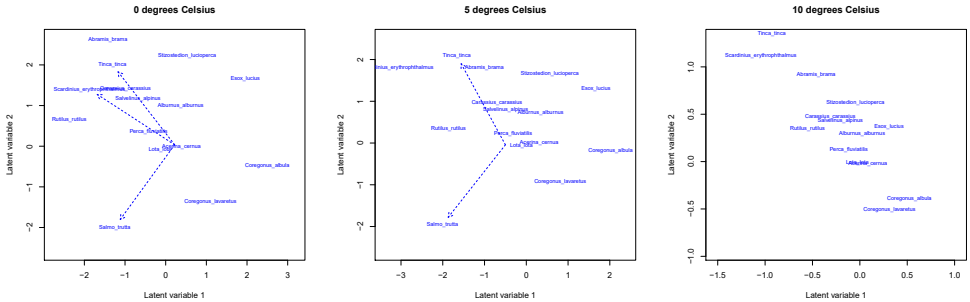


Figure 1: Three figures of the freshwater community in the second ecological application of the GLLVM framework in this thesis. Species associations are modelled as a function of temperature, so that an ordination plot of species can be made that visualizes specis associations at each point along the gradient.

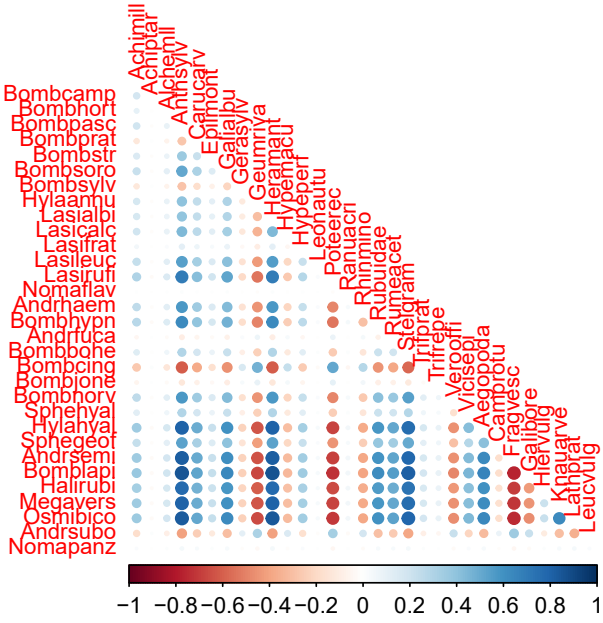


Figure 2: Residual correlations of bee species and plants (abbreviated names, see appendix in 2.2), for the model fitted to the data from the Beediverse project, as part of the second ecological application in this thesis. Bee species are shown on the y-axis and plants on the x-axis

Ultimately, model-based ordination or JSDMs can be understood as the same type of hierarchical model. Which type of visualization is a matter of taste, or "academic upbringing", I suppose. In these two examples, species and sites were hierarchically modelled as a function of environmental predictors. Of course, it is possible to take this a step further and facilitate e.g., including spatial coordinates, functional traits or phylogenetic relationships.

These are the overarching topics and ideas from this thesis that are reviewed in the the third and final section, which includes two review articles. The first article titled *"Next generation ordination with Generalized Linear Latent Variable Models"* relates the newly developed methods to various classical ordination methods. Such a comparison has the potential to make it more straightforward for ecologists to understand model-based ordination, and thus to lower the threshold for transferring to the GLLVM framework for the statistical analysis of data on ecological communities. The final article titled *"Hierarchical Ordination, A unifying framework for drivers of community processes"* in this section and thesis pitches the idea of a new type of framework for ordination and joint species distribution modelling. It encompasses the idea of ecological processes ultimately being represented as a function of the site scores and species loadings in a model. These are then modelled hierarchically following ecological expectations: that space is a site-related aspect that species respond to on a site-by-site basis. Similarly, that species relatedness is a species-specific property.

My personal experience is that ordination is used by field ecologists who love collecting data, whereas SDMs are used by ecologists that are interested in statistical analysis (OK, admittedly that is a bit of a general statement). Would a field resulting from the unification of those two perspectives not be something great? I suppose it's called community ecology.

Background

Ecologists like to collect data, loads of it. Especially in community ecology, researchers often enjoy an elaborate taxonomic knowledge of the species in their communities, so that they go out in the field for days on end to register species that occur under a range of different circumstances. Then, after the dataset has been collected, starts the daunting task of attempting to perform statistical analysis in order to understand the process that underlies the composition of a community with many species.

Ecology is united through a common goal: assessment of the impacts of human-made disturbances and climate change, in order to preserve patterns of biodiversity. It is after all biodiversity that supports societal needs through providing resilience to the ecosystems that provide crucial services (Haines-Young, Potschin, et al. 2010). Statistical ecologists and ecological statisticians support this goal by developing (understanding) of complex models.

Recent years have seen a push for statistical models with increasing complexity to ensure the accuracy of the assessments by ecologists. This includes the use of Generalized Additive Models (GAMs, Wood 2017) and Generalized Linear Mixed-effects Models (GLMMs, Bolker et al. 2009). But, only recently have multispecies models emerged as extension of the GLMM framework, due to the increased need to capture processes that represent a whole ecological community. These multispecies models are aptly named Joint Species Distribution Models (JSDMs, Pollock et al. 2014; Clark et al. 2014). Only after noticing that the large datasets of community ecologists provide a computational issue for JSDMs, was the Generalized Linear Latent Variable Model framework introduced (Warton, Blanchet, et al. 2015).

The GLLVM framework now unites the statistical toolsets of community ecologists. Where multivariate analysis and statistical models for the analysis of species distributions were previously worlds apart, they are now united in a single statistical framework. This thesis further develops the GLLVM framework for application in ecology, by providing new models for classical ecological problems, by demonstrating their application to ecological datasets, and by relating them to methods ecologists are more acquainted with.

After all, ecologists have applied multispecies methods for decades. Almost every community ecologist is familiar with Principal Component Analysis or Correspondence Analysis (Pearson 1901; Hill 1973) and in contrast to GLMMs, ordination methods have the benefit of being easy to use. They also allow for straightforward visual inspection compared to GLMMs, that

require at least a basic understanding of maximum likelihood theory and hypothesis testing in order to draw inference.

However, classical ordination methods make assumptions that are difficult or impossible to verify (ter Braak 1985; Warton, Wright, and Wang 2012). This increases the potential for poor or wrong inference, with potentially detrimental consequences. One of the main benefits of the GL(M)M framework is that assumptions can be checked with the use of residual diagnostics. GLLVMs allow for both: straightforward visual inspection of the model **and** checking assumptions. However, they also inherit the steep learning curve from GLMMs.

For decades classical ordination methods have been interpreted as latent variable models, so that the term "ordination axis" has become synonymous to "latent variable". The word "latent" is defined as "existing, but not yet very noticeable" (Autores 2008), in essence: missing. In statistical models, something that is missing should not occur in the likelihood, so that it needs to be integrated out. This is exactly what GLLVMs do, and it is this process of integration that render GLLVMs much more complex and computationally intensive than classical ordination methods, which treat latent variables as fixed effects instead (Walker and Jackson 2011).

Hawinkel et al. (2019) argue: "if statistical inference were the goal, then random effects would be preferred". They developed a framework for ordination that treats the ordination axes as fixed effects for heuristic reasons, and state: "This (treating the latent variables as random) renders the fitting procedure computationally intensive, without providing a clear improvement to the ordination plot as compared to fixed effects models". This is an arguable choice, as fitting speed should instead be a reason to push for advances in the fitting of GLLVMs, such as with as in Popovic et al. (2019), and not as an excuse to resort to fixed effects ordination methods.

In ecology, latent variables are understood as "gradients". Ecological gradients represent the environment that underlies many of the processes that generate differences in the species composition of ecological communities. In addition to registering species, ecologists often measure the environment as well, which they want to include in their statistical analysis to improve clarity of the patterns that they see. In order to better understand environmental drivers of community composition, constrained ordination methods such as Canonical Correspondence Analysis (ter Braak 1986) were developed to directly relate the environment to the latent variable.

This is where the "latent" term gets confusing, as in constrained ordination methods the latent variable is represented as a function of measured predictors. Hawinkel et al. (2019) writes that

a constrained ordination method for latent variable models is missing, where they supposedly mean "random effect" when they write "latent variable". When all predictors are measured, the meaning "latent variable" is not synonymous to the statistical term of "latent variable" (i.e. a random effect). Then, the latent variables can be considered to be observed, so that it is statistically accurate to treat the parameters of latent variables as a fixed effect instead, so considering latent variable to be synonymous with random effect can be misleading. The same principle applies to the constrained ordination methods developed by Yee (2004), thus concluding that model-based ordination methods with constrained latent variables has been available for decades.

It is the random effects formulation of GLLVMs that connect ordination and species distribution modelling. For decades, these two angles (JSDMs and ordination) have been considered distinctly different. There is a lot ecologists can learn from each other by crossing over to the other world, or by borrowing ecological theory and experience. This is a message that can be found in each article in this thesis. Ultimately, this thesis represents a sorely needed update for the toolset of community ecologists using random effects, in the GLLVM framework.

This year, it is 120 years since Pearson (1901) developed Principal Component Analysis. Even though the deficiencies of that method are well known, it is still being applied by community ecologists. Similarly, Correspondence Analysis and Detrended Correspondence Analysis are still being applied. The GLLVM framework has the potential to replace all of these methods for the analysis of multivariate data in ecology. Hopefully, this thesis will further spark the interest of ecologists in modern methods for the statistical analysis of ecological communities, so that classical ordination methods can eventually be retired.

After all, no one uses linear regression on log-transformed count data anymore either.

References

Autores, V. (Apr. 18, 2008). *Oxford Student's Dictionary*. New Ed edition. Oxford: Oxford University Press. 806 pp.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017). "Variational Inference: A Review for Statisticians". In: *Journal of the American statistical Association* 112.518, pp. 859–877.

Bolker, B. M. et al. (Mar. 1, 2009). "Generalized Linear Mixed Models: A Practical Guide for Ecology and Evolution". In: *Trends in Ecology & Evolution* 24.3, pp. 127–135.

Clark, J. S. et al. (2014). "More than the Sum of the Parts: Forest Climate Response from Joint Species Distribution Models". In: *Ecological Applications* 24.5, pp. 990–999.

Dunn, P. K. and G. K. Smyth (1996). "Randomized Quantile Residuals". In: *Journal of Computational and Graphical Statistics* 5.3, pp. 236–244. JSTOR: 1390802.

Haines-Young, R., M. Potschin, et al. (2010). "The Links between Biodiversity, Ecosystem Services and Human Well-Being". In: *Ecosystem Ecology: a new synthesis* 1, pp. 110–139.

Hawinkel, S. et al. (Feb. 13, 2019). "A Unified Framework for Unconstrained and Constrained Ordination of Microbiome Read Count Data". In: *PLOS ONE* 14.2, e0205474.

Hill, M. O. (1973). "Reciprocal Averaging: An Eigenvector Method of Ordination". In: *Journal of Ecology* 61.1, pp. 237–249. JSTOR: 2258931.

Hui, F. K. C. et al. (Jan. 2, 2017). "Variational Approximations for Generalized Linear Latent Variable Models". In: *Journal of Computational and Graphical Statistics* 26.1, pp. 35–43.

Korhonen, P. (2020). "Fitting Generalized Linear Latent Variable Models Using the Method of Extended Variational Approximation". In.

Niku, J., W. Brooks, R. Herliansyah, F. K. C. Hui, S. Taskinen, and D. I. Warton (May 1, 2019). "Efficient Estimation of Generalized Linear Latent Variable Models". In: *PLOS ONE* 14.5, e0216129.

Niku, J., W. Brooks, R. Herliansyah, F. K. C. Hui, S. Taskinen, D. I. Warton, and B. van der Veen (2021). *Gllvm: Generalized Linear Latent Variable Models*. manual.

Niku, J., F. K. C. Hui, et al. (2019). "Gllvm: Fast Analysis of Multivariate Abundance Data with Generalized Linear Latent Variable Models in r". In: *Methods in Ecology and Evolution* 10.12, pp. 2173–2182.

Niku, J., D. I. Warton, et al. (Dec. 1, 2017). "Generalized Linear Latent Variable Models for Multivariate Count and Biomass Data in Ecology". In: *JABES* 22.4, pp. 498–522.

Pearson, K. (Nov. 1, 1901). "LIII. On Lines and Planes of Closest Fit to Systems of Points in Space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11, pp. 559–572.

Peng, F.-J. et al. (Feb. 1, 2021). "Double Constrained Ordination for Assessing Biological Trait Responses to Multiple Stressors: A Case Study with Benthic Macroinvertebrate Communities". In: *Science of The Total Environment* 754, p. 142171.

Pollock, L. J. et al. (2014). "Understanding Co-Occurrence by Modelling Species Simultaneously with a Joint Species Distribution Model (JSDM)". In: *Methods in Ecology and Evolution* 5.5, pp. 397–406.

Popovic, G. C. et al. (2019). "Untangling Direct Species Associations from Indirect Mediator Species Effects with Graphical Models". In: *Methods in Ecology and Evolution* 10.9, pp. 1571–1583.

Rao, C. R. (1964). "The Use and Interpretation of Principal Component Analysis in Applied Research". In: *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 26.4, pp. 329–358. JSTOR: 25049339.

Ter Braak, C. J. (1985). "Correspondence Analysis of Incidence and Abundance Data: Properties in Terms of a Unimodal Response Model". In: *Biometrics* 41.4, pp. 859–873. JSTOR: 2530959.

– (1986). "Canonical Correspondence Analysis: A New Eigenvector Technique for Multivariate Direct Gradient Analysis". In: *Ecology* 67.5, pp. 1167–1179.

Van der Aart, P. and N. Smeek-Enserink (1974). "Correlations between Distributions of Hunting Spiders (Lycosidae, Ctenidae) and Environmental Characteristics in a Dune Area". In: *Netherlands Journal of Zoology* 25.1, pp. 1–45.

Van der Veen, B. (2021a). *Dyke Data from: De Lange (1972)*. figshare.

– (2021b). *Hunting Spider Data from: Van Der Aart and Smeenk-Enserink (1975)*. figshare.

Walker, S. C. and D. A. Jackson (2011). "Random-Effects Ordination: Describing and Predicting Multivariate Correlations and Co-Occurrences". In: *Ecological Monographs* 81.4, pp. 635–663.

Warton, D. I., F. G. Blanchet, et al. (Dec. 2015). "So Many Variables: Joint Modeling in Community Ecology". In: *Trends Ecol. Evol. (Amst.)* 30.12, pp. 766–779. pmid: 26519235.

Warton, D. I., S. T. Wright, and Y. Wang (2012). "Distance-Based Multivariate Analyses Confound Location and Dispersion Effects". In: *Methods in Ecology and Evolution* 3.1, pp. 89–101.

Wei, T. and V. Simko (2021). *R Package "Corrplot": Visualization of a Correlation Matrix*. manual.

Wood, S. N. (May 18, 2017). *Generalized Additive Models: An Introduction with R, Second Edition*. CRC Press. 497 pp.

Yee, T. W. (2004). "A New Technique for Maximum-Likelihood Canonical Gaussian Ordination". In: *Ecological Monographs* 74.4, pp. 685–701.

The `gllvm` R-package

Throughout this thesis, the `gllvm` R-package has been used for development and dissemination of the models (Niku, Brooks, Herliansyah, Hui, Taskinen, Warton, and van der Veen 2021), especially in the first three articles and one of the applications. Niku, Hui, et al. (2019) developed the original version of the R-package, which intends to make GLLVMs more accessible and easier to fit for ecologists.

GLLVMs are computationally intensive statistical models that can take long to fit, are sensitive to the initial values, and have a steep learning curve. The R-package addresses each of these issues, with the use of approximate methods for integration (Niku, Warton, et al. 2017; Hui et al. 2017), smart generation of initial values (Niku, Brooks, Herliansyah, Hui, Taskinen, and Warton 2019), and by providing a general toolset to examine the results of fitted models (such as creation of a biplot or triplot, tools to construct correlation plots of species associations, and model-selection tools).

Data for examples

In order to perform a limited demonstration of the current functionalities of `gllvm`, I here use two datasets: the hunting spider dataset (also used in the second article of the first section) and the dyke dataset. They were provided by Cajo ter Braak from the example data sets of CANOCO (ter Braak and Smilauer 2012), and have been published, with permission, on figshare (van der Veen 2021a; van der Veen 2021b).

The hunting spider dataset includes 12 species of hunting spiders at 100 sites, which were captured using pitfall traps (see van der Aart and Smeek-Enserink 1974, for details) and includes additional measurements of the environment at 28 sites.

The dyke vegetation data was also used for demonstration by ter Braak (1986). Detailed information on the dataset is available in Canoco, or in the aforementioned article, but it includes binary responses of 125 macrophyte species at 133 sites in Dutch dykes, with measurements of electrical conductivity, phosphate and chloride content of the water, and soil type (either clay, peat or sand). Here, Phosphates was log-transformed as in ter Braak (1986). I use the model.matrix($\cdot$) function to create dummy predictors for the soil type variable (where ($\cdot$) is a placeholder for whatever object we put into a function).

The quadratic response model

In this example I demonstrate how to implement the quadratic response model, as developed in the first article in the first section of this thesis, with the hunting spider data. Quadratic curves are frequently occurring in community ecology, specifically to describe the response of species to the environment. When one has measured predictor variables, a quadratic function can straightforwardly be included in a regression in R using the poly$(\cdot, 2)$ function. However, in a GLLVM, latent variables are included that can represent unmeasured predictors. As such, one might wish to test if species respond to those unknown predictors too. This is similar to the theory behind other ordination methods, such as Correspondence Analysis (ter Braak 1985), which has been one of the key drivers for the popularity of ordination in ecology.

The unique thing about the quadratic response model, is that specifying a quadratic term for each species separately, coincides with the assumption that species have their own unique tolerances to the environment. A more simple model, would be to assume that species have the same tolerances, in essence that all species are a generalist or specialist to the same degree. This can be done using a linear response model, with random row-effects. Here, I will demonstrate how to fit all three models, and then pick the best using information criteria.

```
ftEqTol <- gllvm(Y, family = "poisson", row.eff = "random", num.lv = 2,
    n.init = 3)
```

The `family` argument specifies the response distribution. Since the dataset consists counts for this example, I choose a Poisson distribution (though alternatively a negative-binomial distribution could be used in case of overdispersion). The `row.eff` argument specifies the type of random intercept, which can be an intercept per row as here, or alternatively could specify grouping of a kind, for example if the dataset included multiple plots per site. The `num.lv` argument specifies the number of latent variables to include in the model, and finally the `n.init` argument specifies the number of times the model should be re-fitted, in order to ensure that an optimal solution has been found (due to sensitivity of the approach to the initial values).

I can then use the ordiplot$(\cdot)$ function to construct an ordination diagram. This function has various options to change the visual representation. The most important argument to elaborate on here, is the `predict.region` argument, which construct prediction ellipses for the sites, to give an impression of which sites are predicted to be similar in the ordination.

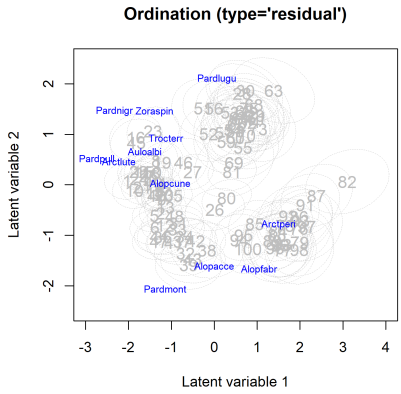


Figure 3: Ordination plot from the equal tolerances model fit to the hunting spider dataset. Grey numbers indicate predicted locations of sites, blue names locations of species.

```
ordiplot(ftEqTol, s.colors = "gray", biplot = TRUE, predict.region = TRUE,
    col.ellips = "gray", lty.ellips = "dashed", alpha = 0.6)
```

Next, I fit a model with the assumption that tolerances are the same for all species, but unique per latent variable, which I will refer to as species common tolerances. I do this using the `quadratic` flag in the gllvm(·) function, which has the options `FALSE`, `LV` (common tolerances), and `TRUE` (unique tolerances for all species).

```
ftComTol <- gllvm(Y, family = "poisson", num.lv = 2, quadratic = "LV",
    n.init = 3)
```

And lastly, I can fit the full quadratic model with the assumption that tolerances are species-specific. Biologically, this model might be most realistic, but it places a heavy burden on the dataset in terms of information required.

```
ftUneqTol <- gllvm(y = Y, num.lv = 2, family = "poisson", quadratic = TRUE,
    n.init = 1, start.struc = "all", starting.val = "zero")
```

As mentioned, GLLVMs are sensitive to the initial values, and with a quadratic response model even more so. As such, the unequal tolerances model by default fits a common tolerances model first, to use as initial values. This option is controlled through the `start.struc` argument in `start.control`. Here, the best set-up (in a maximum likelihood sense) was given with

initial values of zero, and without first fitting a common tolerances model. The options for `starting.val` are `zero`, `res`, and `random`.

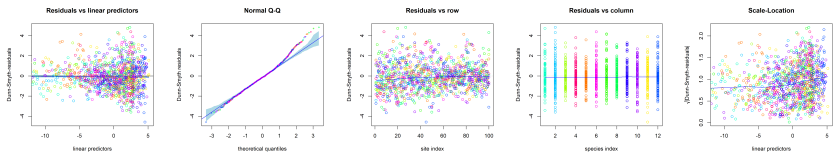
Now, I can use information criteria to determine which of the models fits the hunting spider data best.

```
AICc(ftEqTol, ftComTol, ftUneqTol)
```

```
## [1] 5986.312 6157.395 5733.425
```

The unequal tolerances model fits best, as measured by AICc. It is still important to check if the distributional assumptions have been met, so it is always curcial to examine plots of residuals:

```
plot(ftUneqTol)
```



The residuals calculated here are randomized quantile residuals, which are the gold standard for models that include random effects (Dunn and Smyth 1996). Examining residual plots can be difficult, but the main thing to remember is that there should not be any visible patterns (or deviations in case of the QQ-plot). All lines should be straight, all dots should be (relatively) randomly distributed. Here, I see that there is some deviation in the second plot, that shows the residuals against the theoretical quantiles of the Poisson distribution, which means that the distribution of the residuals is more positively skewed relative to the Poisson. This means, that the model does not *perfectly* represent the data, in that I might have omitted important terms from the model, or I might have to assume a different response distribution (such as the negative-binomial distribution). For demonstration purposes, I will continue to examine the results anyway!

Species optima and tolerances, and their approximate standard errors, can be extracted from the model using the tolerances(·) and optima(·) functions.

The variance explained, which can be used to e.g. decompose variation across latent variables or terms in the model, can be determined using the getResidualCov(·) function:

```
# Residual variance per latent variable for the linear term
getResidualCov(ftUneqTol)$var.q
```

```
##       LV1       LV2
## 100.8808 142.4957
```

```
# for the quadratic term
getResidualCov(ftUneqTol)$var.q2
```

```
##     LV1^2     LV2^2
## 142.96587  30.83365
```

The ordiplot(·) function is used to construct an ordination diagram (here of species optima). However, since species optima can be quite large if they are unobserved, or if too little information is present in the data, creating a nice figure can be challenging. One attempt to improve readability of the species optima in a figure is to point an arrow in their general direction, if species optima are "unobserved": outside of the range of the predicted site scores. Alternatively, a combination of the getResidualCor(·) function and the corrplot function in the R-package with the same name can be used for visualization (Wei and Simko 2021).

```
ordiplot(ftUneqTol, biplot = TRUE, spp.arrows = TRUE, alpha = 0.6, s.colors = "gray")
cormat <- getResidualCor(ftUneqTol)
corrplot::corrplot(cormat, type = "lower", diag = FALSE, order = "AOE",
    mar = c(0, 0, 5, 0), tl.cex = 0.8, addgrid.col = NA)
```

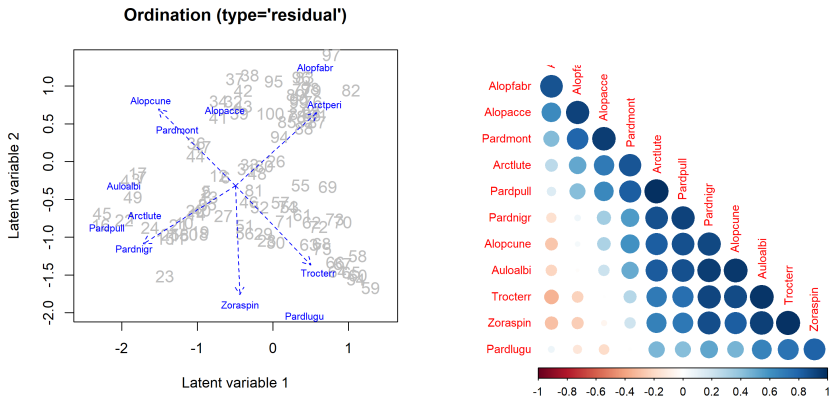


Figure 4: Ordination plot and residual correlations from the unequal tolerances model fitted to the hunting spider dataset.

Ordination with predictors

Until recently, the `gllvm` R-package only supported unconstrained ordination. As such, when including predictor variables in the model, the interpretation of the ordination would shift to that of a residual ordination, i.e. an ordination that is conditional on the predictors, but that does not include the effects of the predictors, some of this is further discussed in either articles in the last section. Here, I will demonstrate this method using the dyke dataset of binary data.

```
# Get the design matrix and scale the predictors
X <- model.matrix(~., X)[, -1]
X <- scale(X)
```

In `gllvm` a multivariate GLM is fitted as:

```
MGLM <- gllvm(Y, X = X, family = "binomial", num.lv = 0)
```

Such a model is also known as a "stacked SDM", since it can also be fitted with independent models per species, since there is a slope per predictor per species, but no terms are shared across species or sites. As such, if the number of predictor variables is large and so is the number of species, including predictors can result in a very large number of parameters to estimate. For data of ecological communities, which can be quite sparse, this is not always a reasonable model to fit. As alternative, ecologists have performed constrained ordination for decades, with methods such as Canonical Correspondence Analysis (ter Braak 1986), or Redundancy Analysis (Rao 1964). In the second article of the first section, we further develop the GLLVM framework for that purpose.

Reduced rank regression is a method akin to multivariate GLMs, where the number of parameters is reduced by the "rank" of the matrix of predictors slopes, which is by default equal to the number of predictors (or the number of species in case that is less). Using reduced rank regression, I fit a model that requires that rank to be equal to a pre-specified number. This then lends a latent variable interpretation due to the way the model is formulated (see the second article in section one for more details). The method is known in ecology under the name of constrained ordination. In constrained ordination, the data is regressed as a function of sites and species, where the sites are again regressed as a function of the predictors. The main difference with classical constrained ordination and the developments in this thesis, is that the former omits the residual of the hierarchical regression, thus assuming that the latent

variable can be represented perfectly by the predictor variables. A step further would be to assume that the residuals of sites are not independent, but instead are spatially autocorrelated, which is a step further than the developments have gone in this thesis, though some discussion on this subjects is available in the final article in this thesis.

A classical ordination can be fitted in the `VGAM` R-package more accurately than with classical ordination methods (i.e. with maximum likelihood; see Yee 2004), or with the implementation that I have also developed for the `gllvm` R-package, which uses numerical optimization (unlike in the `VGAM` R-package). The number of latent variables can be determined by cross-validation, or alternatively, using information criteria (I will use the latter because it is easy!). The code for this in the `gllvm` R-package, for an arbitrary choice of two latent variables, is:

```
RRGLM <- gllvm(Y, X = X, family = "binomial", num.RR = 2)
```

but unlike in other R-packages, it is now possible to formulate a constrained ordination with residual, or with additional random effects in general. Since, let's face it, how often can we be 100% confident that all relevant predictors have been measured, so that there is no residual?! Thus, I assume that I can partially inform the model of what the latent variable is, using predictors, and that there is an additional part I have only information on in the form of species responses, which is how the residual can be understood. The R-code for this is:

```
CGLLVM <- gllvm(Y, X = X, family = "binomial", num.lv.c = 2, method = "LA",
    starting.val = "res", optimizer = "nlminb", n.init = 5)
```

Here, the `optimizer` argument specifies which optimizer the model should be fitted with (currently the possibilities are `optim` or `nlminb`). Changing optimizers can result in a different model fit, or better convergence, similar as in ordinary mixed effects models. So, ideally we want to try different initial values and both optimizers to find the optimal model fit for GLLVMs.

The number of "fully observed" latent variables (i.e. without residual), "partially observed" latent variables (with residual), and unconstrained (or "residual" i.e. completely unmeasured) latent variables can be freely combined using the `num.RR`, `num.lv.c` and `num.lv` arguments (but caution is necessary to prevent overfitting).

Specifying the `method` argument allows us to influence whether the model should be fitted using the Laplace approximation, or using Variational Approximations (the latter is the default).

I chose the Laplace approximation here, since the dataset includes presence-absences. The Variational Approximation implemented for binary responses in the `gllvm` `R`-package has the tendency to underestimate the variance of the latent variables, especially for the probit formulation used here, which is a known deficiency of the method (Blei, Kucukelbir, and McAuliffe 2017), which was especially clear for this dataset (as VA estimated the scale of both LVs to be zero and LA as non-zero). Recent developments of Extended Variational Approximations might offer an alternative solution in that regard (Korhonen 2020). Thus, changing to the Laplace approximation is a good solution here, though fitting models with the Laplace approximation tends to be slower than the Variational Approximation (Niku, Brooks, Herliansyah, Hui, Taskinen, and Warton 2019).

The reduced rank slopes (also known as canonical coefficients in e.g., CCA or RDA) are available under `RRGLM$params$LvXcoef` or can be retrieved with the coef(·) function, or with the summary(·) function (see next page), and the same for the standard deviations of the latent variables.

```
summary(CGLLVM)
```

```
##
## Call:
## gllvm(y = Y, X = X, num.lv.c = 2, num.RR = 0, family = "binomial",
##     method = "LA", starting.val = "res", optimizer = "nlminb",
##     n.init = 5)
##
## Family:  binomial
##
## AIC:  8490.816 AICc:  8511.397 BIC:  11640.03 LL:  -3837 df:  408
##
## Informed LVs:  2
## Constrained LVs:  0
## Unconstrained LVs:  0
## Standard deviation of LVs:  0.3756 0.4471
##
## Formula:  ~ 1
## LV formula:  ~EC + Phosphates + Chlorides + Soil.TypePeat + Soil.TypeSand
##
## Coefficients LV predictors:
##                       Estimate Std. Error z value Pr(>|z|)
## EC(LV1)              0.1643000  0.0020912  78.569  < 2e-16 ***
## Phosphates(LV1)      0.0721775  0.0008617  83.765  < 2e-16 ***
## Chlorides(LV1)       0.1238616  0.0076877  16.112  < 2e-16 ***
## Soil.TypePeat(LV1) -0.1664809  0.0042473 -39.197  < 2e-16 ***
## Soil.TypeSand(LV1)  0.0032786  0.0010693   3.066  0.00217 **
## EC(LV2)              0.1341041  0.0075517  17.758  < 2e-16 ***
## Phosphates(LV2)      0.3118459  0.0019600 159.102  < 2e-16 ***
## Chlorides(LV2)      -0.0991805  0.0021477 -46.180  < 2e-16 ***
## Soil.TypePeat(LV2)  0.0718595  0.0233883   3.072  0.00212 **
## Soil.TypeSand(LV2)  0.1168449  0.0122445   9.543  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The standard errors are all close to zero, which generally means that the model is overfitting, i.e. the model is too complex for the dataset at hand. This is especially likely to happen with small datasets, that contain little information. Again, I will continue here for demonstration purposes.

The summary(·) function by default provides the estimates, standard errors, z-values (wald-statistic) and p-values for the reduced rank predictor slopes. Note, that these do not necessarily correspond with the arrows in an ordination plot, since there the latent variables are by default rotated so that the first latent variable explains maximum variation (and the second, third, etc. thereafter). The `principal` argument in the summary function can be used to retrieve rotated coefficients, the `Lvcoefs` argument to retrieve the loadings, and there are various other arguments. When rotating the slopes, they should not be interpreted with respect to the responses, but relative to the ordination instead (i.e. "chlorides was significantly related to the first latent variable"). Since the models are sensitive to initial values (and also the calculation of the standard errors, z-values and p-values as a consequence), it is important to re-fit each model multiple times to ensure that a stable solution has been found. This might be one of the main downsides of the methods implemented in `gllvm`.

Note, that if the standard deviation of either LVs would be zero, we should switch from `num.lv.c` to `num.RR` for those latent variables, as it is indicative of a more complex model than supported by the data. Fortunately, the scale of both latent variables is estimated as non-zero here.

I can now also use the anova(·) function in the package to do hypothesis testing for the predictors. This function makes use of the well known result that the likelihood ratio test statistic is asymptotically χ^2-distributed with degrees of freedom equal to the difference in the parameters of these two models (so here: one parameters per latent variable and predictor). Then, I can calculate a p-value from from test statistic using the χ^2 probability density function. Note, that the models need to be nested for this, as they are here (the second model is a simpler version of the first). For example, I can refit the model excluding the predictor "Chlorides", by specifying the `lv.formula` argument:

```
CGLLVM2 <- gllvm(Y, X = X, family = "binomial", num.lv.c = 2, method = "LA",
    lv.formula = ~EC + Phosphates + Soil.TypePeat + Soil.TypeSand)
```

```
##   Resid.Df        D Df.diff     P.value
```

```
## 1    16219  0.00000        0
## 2    16217 20.33222        2 3.84517e-05
```

```
anova(CGLLVM,CGLLVM2)
```

It is also possible to combine those arguments with full rank predictors. If combining constrained ordination, with additional predictors, the formula interface has to be used:

```
PCGLLVM <- gllvm(Y, X = X, family = "binomial", num.lv.c = 2, lv.formula = ~EC +
    Phosphates + Chlorides, formula = ~Soil.Type)
```

where `lv.formula` is the formula for the constrained ordination, and `X.formula` is the formula which informs the model which predictors should be modelled in full-rank. Note, that those two formulas cannot include the same predictor variables, and all predictor variables should be provided in the `X` argument. In essence, this performs a partial constrained ordination with latent variables.

Though I did not do so here, information criteria can be used to determine the correct number of reduced ranks, or in general the correct number of constrained and unconstrained latent variables. My recommendation is not to perform model-selection on the included predictor variables, but to mostly focus on the rank (if this causes convergence issues, first scale and centre predictors, and if that does not help perform model-selection on the predictors).

Finally, all the other tools in the `gllvm` R-package can be used for inference with these models, such as creating an ordination diagram with arrows. The arrows that show as less intense red (pink), are predictors of which the confidence interval for the slope includes zero, for at least one of the two plotted dimensions. There are various arguments included in the function to improve readability of the figure, have a look at its documentation. The arrows are always proportional to the size of the plot, so that the predictor with the largest slope estimate is the largest arrow. If the predictors have no effect, the reduced rank slopes will be close to zero.

Although I could fit all models and compare with information criteria and visualize the model that fits best (which is probably what I should do), instead I let you enjoy this beautiful ordination diagram, with prediction ellipses and arrows that indicate the predictor effects, with the note that some additional scaling has been performed to ensure a nice visualization.

```
ordiplot(CGLLVM, biplot = TRUE, xlim = c(-2.5, 2), ylim = c(-3, 3), alpha = 0.202,
    arrow.scale = 0.7, s.colors = "gray", predict.region = TRUE, lty.ellipses = "dashed",
    col.ellipses = "gray")
```

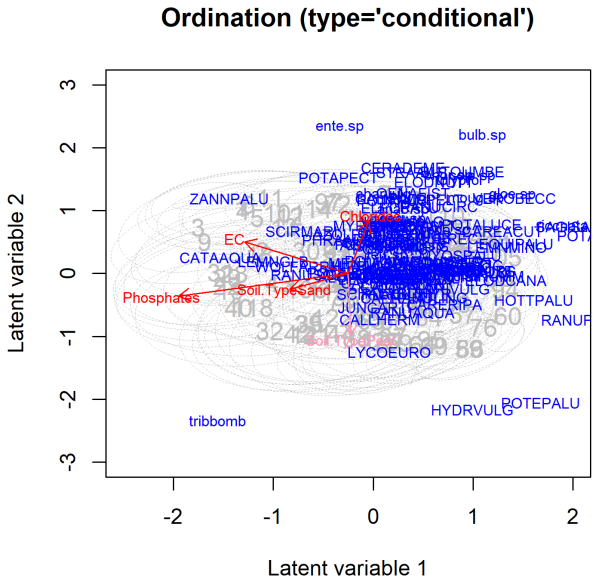


Figure 5: Ordination plot from a model-based ordination with constrained latent variables fitted to the dyke data.

Now, since I have assumed that the latent variables are only partially observed, i.e. they are a function of predictors and random effects, there are three types of distinguishable site scores in the model: 1) "conditional", 2) "marginal", and 3) "residual". The first set contains all effects, the second set contains only fixed effects (i.e. predictor effects) and the third only includes random effects, i.e. the configuration of the sites that is not explained by the predictors. Naturally, if we fitted a model with only the `num.RR` argument, only marginal scores would be available, and only residual scores if we had used the `num.lv` argument.

```
ordiplot(CGLLVM, type = "conditional")
ordiplot(CGLLVM, type = "marginal")
ordiplot(CGLLVM, type = "residual")
```

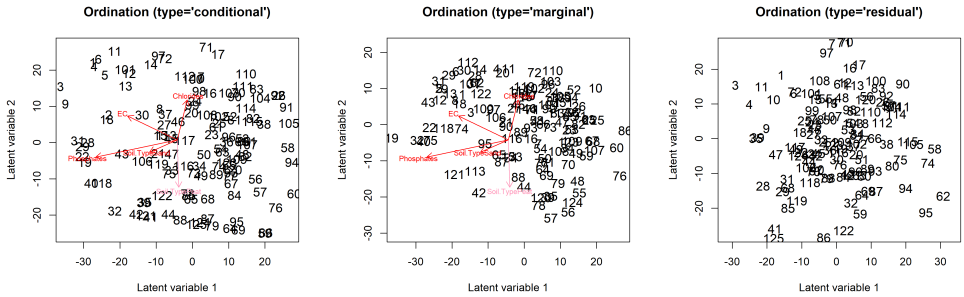


Figure 6: Ordination plots of sites for different components of the latent variables.

Finally, though the discussion in this chapter has mostly been ordination-based, it is helpful to note that in the perspective from the GLLVM framework, JSDMs and ordination are the same modelling approaches. For example, for both datasets and models used here I can visualize the residual correlation matrix to get an impression of species associations. The getResidualCor(·) function in combination with the `corrplot` R-package allows for such visualization.

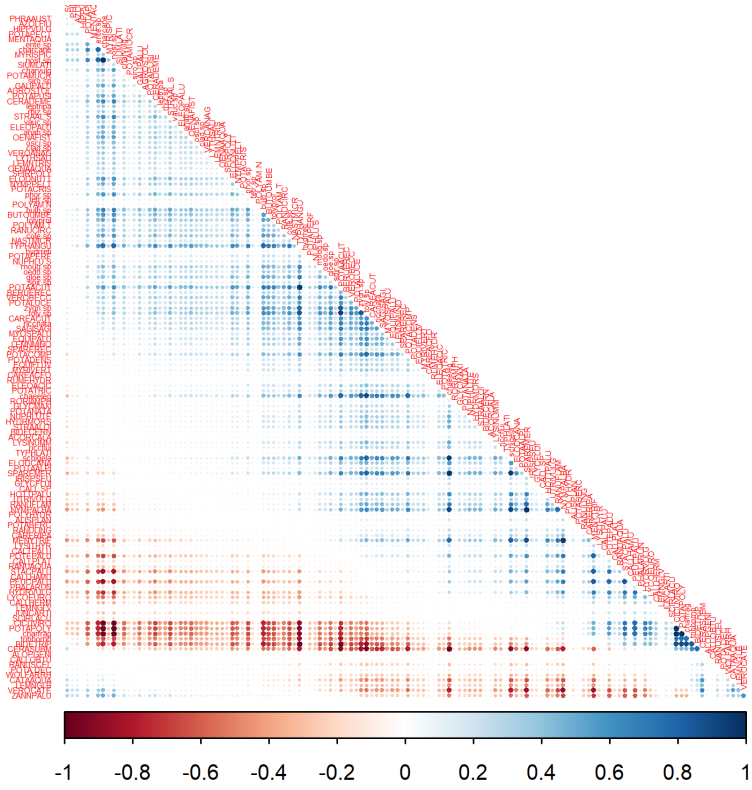


Figure 7: Residual correlations from the dyke dataset and model.

```
cormat <- getResidualCor(CGLLVM)
corrplot::corrplot(cormat, type = "lower", diag = FALSE, order = "AOE",
    addgrid.col = NA)
```

Clearly, for a model with many species, it is difficult to visualize residual correlations in a readable way. Ordination plots have a similar problem, and the solution is to filter species by e.g. functional groups.

In the correlation plot, blue means that two species have a positive relationship in terms of co-occurrence, and red a negative relationship. The more intense the color, the stronger the relationship is predicted to be. Ordination plots show the same information for unconstrained ordination, so that which plot to use is a matter of taste. However, in the case of constrained ordination methods or models using predictors in general, the residual correlation matrix only includes information of species associations not due to the predictors, whereas the ordination plot can include both sources of information. Similarly, residual correlation plots can be difficult to interpret or visualize for a larger number of species.

The `gllvm` R-package provides an important contribution as such, as it makes these complex models faster, and more straightforward, to fit and interpret. Some of the later perspectives in this thesis have not yet been implemented as technical developments are still lacking. Specifically, the hierarchical ordination framework (also known as double constrained ordination e.g. as in Peng et al. 2021), requires implementing (structured) random slope models and the ability to relate traits to the latent variables. One of my main conclusions from this PhD, is that the `gllvm` framework has a lot to offer for ecology, and not nearly all possibilities for development and application have been exhausted just yet.

Contents

1

Dragging community ecology into the 21st century

(with GLLVMs)

1.1 Model-based ordination for species with unequal niche widths

Received: 2 October 2020 | Accepted: 1 March 2021

DOI: 10.1111/2041-210X.13595

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

Model-based ordination for species with unequal niche widths

Bert van der Veen[1,2,3] | Francis K. C. Hui[4] | Knut A. Hovstad[3,5] | Erik B. Solbu[1] | Robert B. O'Hara[2,3]

[1]Department of Landscape and Biodiversity, Norwegian Institute of Bioeconomy Research, Trondheim, Norway

[2]Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

[3]Centre of Biodiversity Dynamics, Department of Biology, Norwegian University of Science and Technology, Trondheim, Norway

[4]Research School of Finance, Actuarial Studies and Statistics, Australian National University, Canberra, ACT, Australia

[5]The Norwegian Biodiversity Information Centre, Trondheim, Norway

Correspondence
Bert van der Veen
Email: bert_van_der_veen@hotmail.com

Funding information
Research Council of Norway, Grant/Award Number: 272408/F40; Australian Research Council

Handling Editor: Javier Palarea-Albaladejo

Abstract

1. It is common practice for ecologists to examine species niches in the study of community composition. The response curve of a species in the fundamental niche is usually assumed to be quadratic. The centre of a quadratic curve represents a species' optimal environmental conditions, and the width its ability to tolerate deviations from the optimum.
2. Most multivariate methods assume species respond linearly to niche axes, or with a quadratic curve that is of equal width for all species. However, it is widely understood that some species have the ability to better tolerate deviations from their optimal environment (generalists) compared to other (specialist) species. Rare species often tolerate a smaller range of environments than more common species, corresponding to a narrow niche.
3. We propose a new method, for ordination and fitting Joint Species Distribution Models, based on Generalized Linear Mixed-effects Models, which relaxes the assumptions of equal tolerances.
4. By explicitly estimating species maxima, and species optima and tolerances per ecological gradient, we can better explore how species relate to each other.

KEYWORDS
joint species distribution model, model-based ordination, niche model, unconstrained quadratic ordination, unimodal response

1 | INTRODUCTION

One of the key topics addressed by community ecology is the exploration of community composition. To that end, species communities are surveyed at locations along environmental gradients. The ecological niche is then reflected in the observed distribution of a species. A species exhibits its maximum abundance, or has the highest probability of occurrence, at the optimum of the niche. The limits of a species distribution correspond to the limits of the niche, controlled by a species' tolerance to a range of environmental conditions. Different species vary in their ability to tolerate deviations from the optimum, reflecting differences in niche width, and indicating different places on the specialist–generalist spectrum.

Correspondence analysis (CA) is often used to estimate the optima of species niches with quadratic response curves. It implicitly approximates the fit of a quadratic model, which functions best under the assumptions of equally spaced optima, sites being well within the range of species optima, equal tolerances and equal or independent maxima (ter Braak, 1985). The combination of assuming equally spaced optima, equal maxima and equal tolerances gives an early niche model called the species packing model (MacArthur & Levins, 1967). The relationship of the species packing model to

CA has added to its popularity among applied ecologists (Wehrden et al., 2009).

Recent advances in the estimation of species niches have focused on performing ordination with explicit statistical models, such as Generalized Linear Latent Variable Models (GLLVMs; Warton et al. 2015). With intercepts included for row standardization, GLLVMs can fit a quadratic response curve, assuming species have equal tolerances (Hui et al., 2015; Jamil & ter Braak, 2013). When predictor variables are included, a GLLVM with quadratic response model partitions species distributions in observed (fixed effects) and latent or unobserved (random effects), similar to the partitioning of fixed and random effects in mixed-effects models when predictors are included.

The GLLVM framework is well known for its capability to fit Joint Species Distribution Models (JSDMs; Ovaskainen et al., 2017; Pollock et al., 2014; Tobler et al., 2019; Zurell et al., 2020). In the context of JSDMs, GLLVMs assume species abundances are correlated due to similarity in response to ecological gradients, modelled with predictor variables and latent variables. Latent variables can be understood as combinations of missing predictors, so that GLLVMs allow us to parsimoniously model species distributions. They are equivalent to ordination axes, representing complex ecological gradients (Halvorsen, 2012). Recently, the use of GLLVMs to perform model-based ordination has increased in popularity (Björk et al., 2018; Damgaard et al., 2020; Inoue et al., 2017; Lacoste et al., 2019; Paul, 2020). However, existing GLLVMs assume that species respond to latent variables linearly, just as all classical ordination methods do (Jamil & ter Braak, 2013). In contrast, it is widely understood that species have unequal tolerances, so that the assumption of linear responses, or at best quadratic responses with equal tolerances, is unlikely to hold in practice.

In this paper, our goal was to overcome the assumptions of equal tolerances, by formulating a GLLVM where species are allowed to respond to the latent variables in a quadratic fashion. To our knowledge, there has been no attempt to implement such a GLLVM until now. The quadratic term allows to fully estimate species niches, so that species optima and tolerances per latent variable and species maxima can all be estimated explicitly. Explicitly estimating the combination of these three parameters gives unique insight into reasons for species rarity, whether it is due to low abundance or probability of occurrence (maxima), a high degree of habitat specialization (tolerance) or due to unsuitable observed environmental conditions (optima). Due to the model-based nature of the proposed ordination method, it is possible to calculate confidence intervals for each set of parameters, providing unparalleled benefits for inference when using ordination. Additionally, assuming a quadratic response model allows to implement the concept of gradient length, as in Detrended Correspondence Analysis (DCA; Hill and Gauch, 1980), which is a measure of beta diversity commonly used by ecologists.

In contrast to classical ordination methods, GLLVMs model the latent variables as unobserved, treating them as random rather than fixed (Walker & Jackson, 2011), which consequently have to be integrated over in the likelihood. Here, we develop a variational approximation (VA) implementation after Hui et al. (2017) and Niku et al. (2019), to perform calculations quickly and efficiently. In addition to presenting the GLLVM with quadratic response model, we perform simulations to evaluate the accuracy of the VA implementation, and the capability of the GLLVM with quadratic response model to retrieve the true species-specific parameters and latent variables. We use two real-world datasets to demonstrate the use and interpretation of the proposed GLLVM with quadratic responses: (a) a small dataset of hunting spiders in a Dutch dune ecosystem (van der Aart & Smeek-Enserink, 1974), and (b) a larger dataset of Swiss alpine plant species on a strong elevation gradient (D'Amen et al., 2018).

2 | MODEL FORMULATION

The ecological niche for each species $j = 1...p$ is described here by a quadratic function involving three parameters: the optimum u_{jq} for latent variable $q = 1...d$ stored in the vector $\boldsymbol{u}_j = \{u_{j1}...u_{jq}\}$, the tolerance t_{jq} for latent variable q stored in the vector $\boldsymbol{t}_j = \{t_{j1}...t_{jq}\}$ and a species' overall maximum c_j. Optima $\boldsymbol{u}_j$ are the locations on the ecological gradients where a species exhibits its highest abundance or probability of occurrence (the maximum c_j). The tolerances $\boldsymbol{t}_j$ are a measure of the width or breadth of the niche, indicating if a species is a generalist or specialist on each ecological gradient.

Consider an $n \times p$ matrix of observations, where y_{ij} denotes the response of species j at sites $i = 1...n$. Then, we assume that conditional on a vector $\boldsymbol{z}_i \sim \mathcal{N}(\boldsymbol{0}, \mathrm{I})$ of d latent variables where $d \ll p$, the responses y_{ij} at site i are independent observations from a distribution whose mean, denoted here as $\mathrm{E}(y_{ij}|\boldsymbol{z}_i)$, is modelled as:

$$\begin{aligned} g\left\{\mathrm{E}\left(y_{ij}|\boldsymbol{z}_i\right)\right\} &= c_j - \sum_{q=1}^{d}\left\{\frac{\left(z_{iq}-u_{jq}\right)^2}{2t_{jq}^2}\right\} \\ &= c_j - \sum_{q=1}^{d}\left(\frac{u_{jq}^2}{2t_{jq}^2} - \frac{z_{iq}^2}{2t_{jq}^2} + \frac{z_{iq}u_{jq}}{t_{jq}^2}\right), \end{aligned} \tag{1}$$

where $g\{\cdot\}$ is a known link function (e.g. the log link when responses are assumed to be Poisson, negative-binomial or gamma distributed, the probit link when the responses are assumed to be Bernoulli or ordinal distributed and the identity link for responses that are assumed to be Gaussian distributed).

For a closer comparison to the GLLVM with linear response model (Hui et al., 2015), we formulate the GLLVM with quadratic species response curves in terms of matrix notation:

$$g\left\{\mathrm{E}\left(y_{ij}|\boldsymbol{z}_i\right)\right\} = \beta_{0j} + \mathbf{z}_i^{\top}\boldsymbol{\gamma}_j - \mathbf{z}_i^{\top}\mathbf{D}_j\mathbf{z}_i, \tag{2}$$

with a species-specific intercept β_{0j} that accounts for species mean abundances, and a vector of coefficients per species for the linear term $\boldsymbol{\gamma}_j$. We can see that a third term is added here to the existing structure of a GLLVM with linear species responses, which models tolerances per species and latent variable. Specifically, we introduce a diagonal matrix $\boldsymbol{D}_j$ of positive-only quadratic coefficients, with each diagonal

element being the quadratic effect for latent variable q and species j. The sign constraint ensures that species exhibit concave quadratic curves only. The proposed model could instead be used to estimate species minima rather than maxima, though we did not do that here as clear ecological foundations for such a model are lacking.

Let D_{jqq} denote the diagonal elements of $\boldsymbol{D}_j$ for latent variable q. Then we are able to derive the following connections between the parameters in Equations (1) and (2): $\beta_{0j} = c_j - \frac{1}{2}\sum_{q=1}^{d} u_{jq}^2/t_{jq}^2$, $\gamma_{jq} = u_{jq}/t_{jq}^2$, and $D_{jqq} = 1/\left(2t_{jq}^2\right)$. Similarly, for the formulation in Equation (2), the parameters in Equation (1) can be retrieved: $c_j = \beta_{0j} + \frac{1}{4}\sum_{q=1}^{d}\gamma_{jq}^2/D_{jqq}$, $u_{jq} = \gamma_{jq}/\left(2D_{jqq}\right)$, and $t_{jq} = 1/\sqrt{2D_{jqq}}$.

Additionally, row intercepts or predictors can be included as in Hui et al. (2017), or species traits as in Niku et al. (2019), though we have chosen to omit those terms here and focus on the case of unconstrained ordination.

Four special cases of the GLLVM with quadratic response model, as formulated in Equation (2), are worth discussing: (a) $\boldsymbol{D}_j = D$, that is, common tolerances for species, (b) $\boldsymbol{D}_j = D_{11}\mathbf{I}_d$ where $\mathbf{I}_d$ is a $d \times d$ identity matrix, that is, equal tolerances for species and latent variables, (c) when $\boldsymbol{D}_j = 0$ for a subset of the p species and (d) when $\boldsymbol{D}_j = 0$ for all p species. The first case assumes tolerances to be the same across species, but not latent variables. This species-common tolerances model might prove useful in practice, as it requires fewer observations per species than when estimating quadratic coefficients for all species, but still explicitly includes quadratic species responses in contrast to the simpler GLLVM with linear responses. In the second case, the quadratic term is not species or latent variable specific, so that it is equivalent to the GLLVM with linear species responses and random row intercepts as presented in Hui et al. (2015), which assumes tolerances to be the same for all species and latent variables. In the third case, some species respond to the latent variable linearly, while others exhibit quadratic responses. The fourth case is the most basic GLLVM with linear responses, which is the current standard in many software packages for JSDMs and model-based ordination, for example, boral (Hui, 2016), HMSC-R (Tikhonov et al., 2021) and gllvm (Niku et al., 2020).

3 | MODEL INTERPRETATION

In this section, we derive and discuss various tools that are commonly used in the application of JSDMs and ordination, such as calculating residual correlations, partitioning or decomposing residual variance, calculating gradient length and visualizing the ordination, and demonstrate how they can be adapted to the proposed GLLVM with quadratic response model.

3.1 | Residual covariance matrix

One aspect of GLLVMs is known for is modelling species residual correlations (Blanchet et al., 2020; Zurell et al., 2018), calculated from the residual covariance matrix. To facilitate calculation of the residual covariance matrix, we can reparameterize all GLLVMs as a multivariate mixed-effects model with a residual term:

$$g\left\{\mathrm{E}\left(y_{ij}|\mathbf{z}_i\right)\right\} = \beta_{0j} + \epsilon_{ij}. \tag{3}$$

Here, ϵ_{ij} accounts for any residual information that is not accounted for by fixed effects in the model, such as predictors or intercepts (Warton et al., 2015). Assuming the latent variables are independent for all sites, the elements of the residual covariance matrix are given by:

$$\Sigma_{jk} = \mathrm{cov}\left(\epsilon_{ij}, \epsilon_{kl}\right), \quad \forall i,k = 1\ldots n, j,l = 1\ldots p.$$

For a length p vector $\boldsymbol{\epsilon}_i$, existing JSDM implementations (e.g. Pichler & Hartig, 2020; Pollock et al., 2014) assume $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, that is, the residual term follows a multivariate normal distribution. For the GLLVM with linear species responses, it is straightforward to show that with $\epsilon_{ij} = \mathbf{z}_i^\top \boldsymbol{\gamma}_j$, then $\boldsymbol{\epsilon}_i \sim \mathcal{N}\left(\mathbf{0}, \Gamma\Gamma^\top\right)$, where $\boldsymbol{\Gamma}$ is a $p \times d$ matrix of species linear coefficients for the latent variables $\boldsymbol{\gamma}_j$. In essence, GLLVMs preform a low rank approximation to the covariance matrix of a residual term. The rank of this residual covariance matrix is equal to the number of estimated latent variables d in the model for the GLLVM with linear species responses.

Turning to the GLLVM with quadratic response model, where $\epsilon_{ij} = \mathbf{z}_i^\top \boldsymbol{\gamma}_j - \mathbf{z}_i^\top \boldsymbol{D}_j \mathbf{z}_i$, the elements of the residual covariance matrix are:

$$\Sigma_{\mathrm{quad},jk} = \sum_{q=1}^{d}\left(\gamma_{jq}\gamma_{kq} + 2D_{jqq}D_{kqq}\right), \tag{4}$$

for which a proof is given in Appendix S1. This can be rewritten in terms of the species optima $\boldsymbol{u}_j$ and tolerances $\boldsymbol{t}_j$:

$$\Sigma_{\mathrm{quad},jk} = \sum_{q=1}^{d}\left\{\left(t_{jq}^2 t_{kq}^2\right)^{-1}\left(0.5 + u_{jq}u_{kq}\right)\right\}. \tag{5}$$

Equations (4) and (5) additionally serve to demonstrate how to partition and decompose the residual variance of the GLLVM with quadratic response model, for example, per latent variable, for the linear and quadratic term separately, or both. Variance partitioning is commonly used in the application of ordination methods, for example, to determine fit (Øland, 1999), or to explore causes of residual variance (Borcard et al., 1992; Øland & Eilertsen, 1994). Predictor variables can be included in the model to account for the residual variance otherwise accounted for by the latent variables. The residual variance can be used to identify indicator species, that is, those species that best represent an ecological gradient, or to calculate a measure of R^2 (Nakagawa & Schielzeth, 2013).

Under the assumption of latent variables with zero mean, the linear and quadratic terms in the model are independent. As such, the rank of the residual covariance matrix is double that of a GLLVM with linear species responses and the same number of latent variables, $2d$. The additional quadratic term thus allows us to account for more

residual correlations between species, with fewer latent variables. This corresponds to the ecological notion that species often respond to few major complex ecological gradients (Halvorsen, 2012). From this, we see that when the number of latent variables in a GLLVM with quadratic response model exceeds ½*p*, there are more parameters included than in a JSDM with an unstructured residual covariance matrix. However, this is not an issue here, since for ordination purposes we are only interested in cases where there are much fewer latent variables *d* than species *p*.

3.2 | Gradient length

The length of an ecological gradient is of great interest to ecologists in the use of ordination, because it is a measure of beta diversity (Oksanen & Tonteri, 1995). Longer gradients indicate higher diversity, as spacing between sites in latent space is potentially larger. In the past, it has been emphasized that short gradients are better analysed using linear ordination methods, and longer with unimodal methods (ter Braak & Prentice, 1988). However, the GLLVM with quadratic response model allows species to exhibit both linear and unimodal responses, and so it is appropriate for both, removing the need to switch between ordination methods as a consequence of (the lack of) unimodal species responses. Regardless, gradient length could be used to decide between response models instead of, for example, information criteria. To determine gradient length from the proposed GLLVM with quadratic response model, we rescale the latent variables $\boldsymbol{z}_i$ with a diagonal covariance matrix $\boldsymbol{G}$ of size $d \times d$, to calculate ecological gradients $\tilde{\boldsymbol{z}}_i$. The measure of gradient length calculated here can be interpreted in the same manner as the gradient length provided by DCA (Hill & Gauch, 1980).

First, for a species-common tolerances model, we note that the quadratic term in Equation (2), that is, $\boldsymbol{z}_i^\top \boldsymbol{D}\boldsymbol{z}_i$, can instead be written as $\sum_{q=1}^{d} z_{iq}^2 D_{qq}$, so that $\tilde{z}_{iq} = z_{iq}\sqrt{D_{qq}}$, and $\tilde{\boldsymbol{z}}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G})$, where $\boldsymbol{G} = 2\boldsymbol{D}$. Then, the length per ecological gradient is approximately $4G_{qq}^{\frac{1}{2}}$ (i.e. the approximate width of a normal distribution).

Second, for the species-specific tolerances model, we note that one of the uses of gradient length in the past has been to rescale the latent variables so that an ordination diagram can be understood in terms of compositional turnover (Hill & Gauch, 1980). This requires the mean species tolerances to be one (as is the case for the species-common tolerances model, under the rescaling suggested above), so that the covariance matrix of the ecological gradient in the species-specific tolerances model is $G_{qq} = \frac{1}{2p}\sum_{j=1}^{p} D_{jqq}$ and the matrix of quadratic coefficients $\boldsymbol{D}_j$ is scaled by the inverse of the covariance matrix of the ecological gradient, $\boldsymbol{G}^{-1}$. However, we choose to use the median of the species tolerances t_{jq} instead, as it more accurately represents gradient length with both linear and quadratic responses of species in the model. In general, the proposed quadratic model allows further exploration of measures of gradient length by, for example, using the mean tolerance of species with clear quadratic responses, rather than the median of all tolerances.

3.3 | Ordination diagram

Usually, results from an ordination are inspected visually, by jointly plotting site and species scores. For a GLLVM with linear responses, this can be done by constructing a biplot (Gabriel, 1971). Biplots perform a linear approximation of a matrix, and thus are expected to perform poorly when species exhibit quadratic responses: biplots will create an arch when the residual variance of the linear term is smaller than the residual variance of the quadratic term. When the linear and quadratic terms are independent, as is the case here (see above), a biplot can visualize them separately.

Instead, we propose that species optima and tolerances can be plotted directly, so that species niches are visualized in a two-dimensional latent (ecological) space from a top-down perspective. However, since species are allowed to exhibit linear responses in the quadratic response model, optima and tolerances can be very large. If plotting both directly, this will lead to species with large optima and wide niches dominating the plot. The first issue can be prevented by only visualizing species optima that are close to, or within, the range of the estimated site scores, and by using arrows to indicate the location of the remaining optima (similarly as in Gabriel, 1971). The widths of the niches can be represented as ellipses using the precision of estimated species tolerances, to provide an impression of species co-occurrence patterns. The precision, calculated as the inverse of the squared species tolerances $1/t_{jq}^2$, can be interpreted as 'narrowness' of the ecological niche (i.e. a small precision corresponds to a wide niche). Then, a larger ellipse corresponds to a larger residual variance of the quadratic term of a latent variable, drawing emphasis to potential indicator species.

Additionally, information on sites, such as the predicted locations and prediction regions, can be added (Hui et al., 2017). Information for the sites can be used to infer the distance of sites to the species optima (i.e. the suitability of sites for species), or to the edges of species niches (see the hunting spiders example below).

Finally, based on the discussion in the two subsections above, there are two ways of scaling the ordination diagram: (a) by the residual variance per latent variable, or (b) by using the mean or median tolerance. In the first scaling, the diagram is scaled to draw attention to the latent variable that explains most variance in the model. However, the second scaling has a more ecological intuitive interpretation; if the tolerances are assumed to be common for species, the second scaling produces an ordination diagram in units of compositional turnover (Gauch, 1982). When the linear term in the model does not explain a larger proportion of the total residual variance per latent variable relative to the quadratic term, these scalings produce similar results.

4 | MODEL ESTIMATION

We propose to use VAs (Hui et al., 2017) for estimation and inference for the GLLVM with quadratic response model. Broadly speaking, VA is a general technique used to provide a closed-form

approximation to the marginal log-likelihood of a model with random effects or latent variables, when an analytical solution is not available. Computationally, VA can be orders of magnitude faster than MCMC, numerical integration or even the Laplace approximation (Niku et al., 2019), and without loss of accuracy (Hui et al., 2017). However, the calculation of the VA log-likelihood needs to be derived on a case-by-case basis. In contrast, the Laplace approximation can be applied automatically in many cases (Kristensen et al., 2016), although it is not possible to apply that here for the GLLVM with quadratic response model (K. Kristensen, pers. comm., 8 March 2019).

The marginal log-likelihood of a GLLVM is given by:

$$\mathscr{L}(\Theta) = \sum_{i=1}^{n} \log \left\{ \int_{-\infty}^{\infty} \prod_{j=1}^{p} f\left(y_{ij} | \boldsymbol{z}_i, \Theta\right) h(\boldsymbol{z}_i) d\boldsymbol{z}_i \right\}, \quad (6)$$

where $f(y_{ij}|\boldsymbol{z}_i, \Theta)$ is the distribution of the species responses given the latent variables. As mentioned previously, and as per Hui et al. (2015), we assume the distribution of the latent variables $h(\boldsymbol{z}_i)$ to be multivariate standard normal, that is, $h(\boldsymbol{z}_i) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. The vector Θ includes all parameters in the model $\Theta = \left\{\beta_{01} \ldots \beta_{0j}, \gamma_{11} \ldots \gamma_{jq}, D_{111} \ldots D_{jqq}\right\}^{\top}$.

In VA, we construct a lower bound to Equation (6), by assuming that the posterior distribution of the latent variables can be approximated by a closed-form distribution, for example, a multivariate normal distribution (this is also referred to as the variational distribution). We then treat this lower bound as our new objective function, on which we base estimation and inference of the model parameters, as well as predictions of the latent variables. More details on the motivation and background of variational approximations are available in the study by Ormerod and Wand (2010, 2012). Hui et al. (2017) showed that, for GLLVMs with linear responses, the optimal variational distribution is multivariate normal $\boldsymbol{z}_i \sim \mathcal{N}\left(\boldsymbol{a}_i, \boldsymbol{A}_i\right)$, with mean $\boldsymbol{a}_i$ and covariance matrix $\boldsymbol{A}_i$, so we will adopt this choice here as well. While we do not anticipate a multivariate normal distribution to be the optimal variational distribution for a GLLVM with quadratic response model, we nevertheless choose to follow the same assumption to facilitate computational efficiency and a closed form for the resulting VA log-likelihood. The means of the variational distribution $\boldsymbol{a}_i$ can be understood as predicted locations of sites, that is, site scores in an ordination. The covariance matrices of the variational distributions $\boldsymbol{A}_i$ provide the necessary information to construct prediction regions.

In Appendix S2 we provide derivations for the log-likelihood of common response types in community ecology, such as count data (Poisson, a Poisson–Gamma derivation of the negative-binomial distribution for overdispersed counts and both assuming a log-link function), binary data and ordinal data (both with probit-link function), as well as positive continuous data (gamma, with log-link function) and continuous data (Gaussian, with an identity-link function). Additionally, some information on calculating approximate confidence intervals for (functions of) the parameters is included in Appendix S2. Recommendations on stabilizing the fitting of GLLVMs with a quadratic response model are included in Appendix S3.

5 | SIMULATION STUDY

To assess how well the proposed model retrieves the true latent variables $\boldsymbol{z}_i$, optima $\boldsymbol{u}_j$ and tolerances $\boldsymbol{t}_j$, we performed simulations for six response distributions; (1) Gaussian, (2) gamma, (3) Poisson, (4) negative-binomial, (5) Bernoulli and (6) ordinal. The `R` code used for the simulations is provided in Appendix S4. For each of the distributions, we simulated 1,000 datasets with different numbers of sites and species. A consequence of restricting the quadratic response model to concave shapes only is that it often simulates a large number of negative values (on the link scale, generally more so than the GLLVM with linear species responses), providing a challenge in testing its accuracy, especially for small datasets.

First, to study the accuracy of the VA approximation, we simulated datasets of p = 20–100 species in increments of 10, while keeping the number of sites constant at n = 100. Hui et al. (2017) argued that the VA log-likelihood is expected to converge to the true likelihood as $p \to \infty$, as with many species the posterior for the site scores is likely to be approximately normal due to the central limit theory. This will allow us to study the finite sample properties of the VA approximation for the proposed model. Second, to explore the sample size required to accurately estimate the species-specific parameters, for example, species optima $\boldsymbol{u}_j$ and tolerances $\boldsymbol{t}_j$, we simulated datasets of n = 20–100 sites in increments of 10, while keeping the number of species constant at p = 100.

As a true model, we considered a GLLVM with quadratic response model and d = 2 latent variables, which was constructed as follows. The latent variables were simulated following a multivariate standard normal distribution, that is, $\boldsymbol{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Second, the species maxima c_j were simulated as Uniform(2,6), as this was approximately the range of species maxima in the best fitting model for the hunting spider dataset below. Next, the true optima u_{jq} were simulated within the range of the realized latent variables (approximately between −2 and 2) following a uniform distribution. Lastly, species tolerances were simulated as Uniform(0.2,1), corresponding to species niches ranging from narrow to the full width of the latent variable. Resulting species-specific intercepts β_{0j} from Equation (2) approximately ranged between −15 and 20, but tended to be more positive than negative, with a median of 2.6. For the Gaussian, negative-binomial and gamma distributions, the dispersion parameter for all species was set equal to 1. For the ordinal distribution, we assumed six classes with the true cut-offs being 0, 1, 2, 3, 4, 5, meaning that species were most often absent (category 0), while they were rarely very abundant (category 5). When fitting a model to each simulated dataset, we assumed the number of latent variables was known prior to fitting (i.e. we did not select the number of latent variables).

We measured performance of the GLLVM with quadratic response model by the prediction of the latent variables $\boldsymbol{z}_i$ and the species optima $\boldsymbol{u}_j$. The species optima are a function of both the

linear and quadratic coefficients and should provide a good overall measure of performance for retrieving the true species-specific parameters, in addition to being of specific interest to ecologists. We measured discrepancy to the true parameter values using the Procrustes error (Peres-Neto & Jackson, 2001). For this, we excluded the optimum of the first species on the second latent variable as this was fixed to zero for reasons of parameter identifiability (Hui et al., 2015). Since the GLLVM with quadratic response model allows species to exhibit linear responses, which have optima tending to infinity, we also chose to remove all optima larger than 10 and smaller than −10, that is, for those species that clearly lacked a sufficiently strong quadratic signal in the simulated datasets. Including these optima would result in a biased view of the accuracy of the optima that can be estimated by the model. For clarity and transparency, we additionally present the number of optima removed for each of the datasets, to further provide an impression of the data requirements of the proposed model.

For all of the models fitted to Gaussian and gamma response datasets, typically none or only a few optima were excluded, meaning that the median number excluded was zero. In general, and not surprisingly, more optima were excluded for models fitted to datasets where n/p was small and for discrete distributions. For example, when $n = 20$ sites and $p = 100$ species, so that the true model included a total 200 species optima, the median number of optima excluded for datasets with Poisson responses was 4 (2–5, first and third quartiles), for datasets with negative-binomial responses this was 7 (5–10), for datasets with Bernoulli responses this was 44 (40–47) and for datasets with ordinal responses this was 20 (17–24). In contrast, for datasets where n/p was large, considerably fewer optima were excluded across all response types. For example, when $n = 100$ and $p = 100$, and for Poisson responses, the median of excluded optima was 1 (1–3), for negative-binomial response datasets this was 6 (5–7), while for Bernoulli response datasets the median number of optima excluded was with a median of 29 (27–32) still large and for ordinal response datasets this was 13 (11–15).

The symmetric Procrustes error per distribution and for the different sized datasets is presented in Figure 1. As expected, the GLLVM with quadratic responses was more accurate for datasets with larger p and larger n. For all distributions, the latent variables

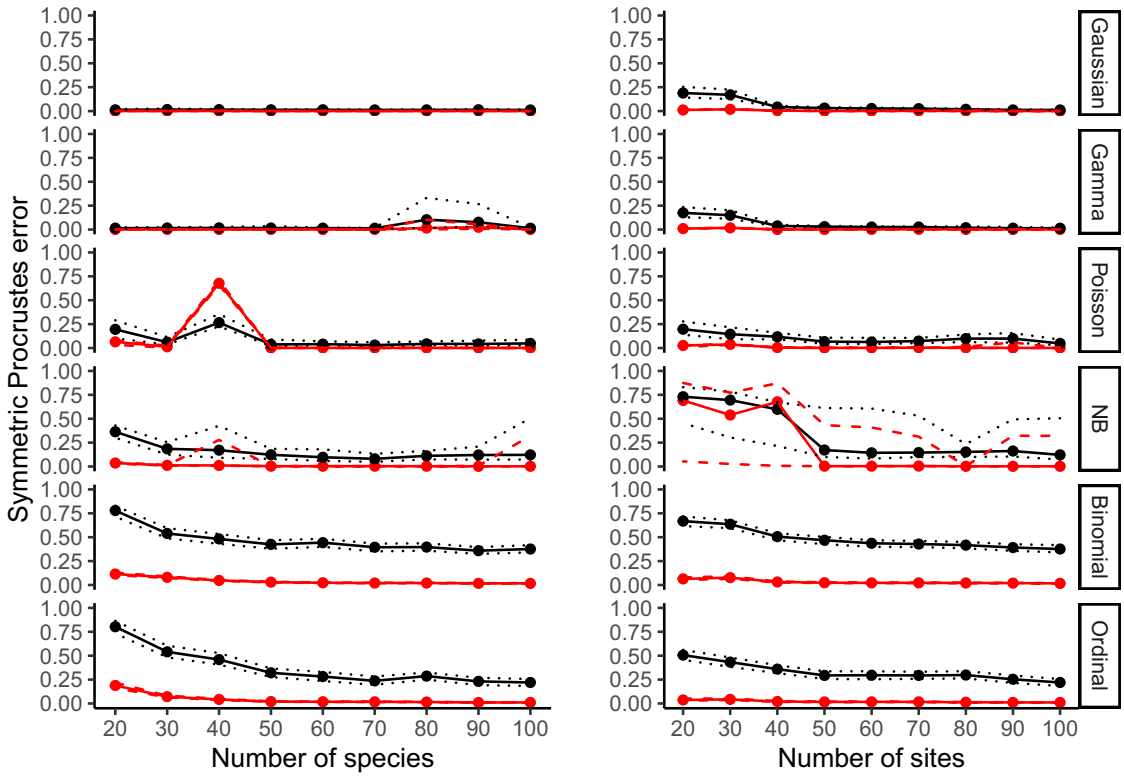


FIGURE 1 Simulation results for the 1,000 GLLVMs fitted to each dataset and response distribution, with the symmetric Procrustes error calculated based on optima that could be estimated (optima outside the range (−10,10) were excluded). The left column shows simulations where the number of sites was kept constant at $n = 100$, and analogous for the right column with $p = 100$. The figure includes the median Procrustes Error for species optima (black) and latent variables (red), with the first and third quartiles represented as dotted (optima) and dashed (latent variables) lines

were often better retrieved than the species optima. This is not surprising, as the species optima are a function of two parameters, particularly the inverse of the quadratic coefficients, so that a small change in the quadratic coefficients can result in a large change in the species optima. When fitted to Gaussian or gamma response datasets, the model performed best. The accuracy of the estimated species optima and latent variables was only slightly lower for datasets with Poisson responses, and was also similar for datasets with negative-binomial responses and a large number of sites. If the number of sites is small, the variation in accuracy of the latent variable and of species optima was considerably larger for datasets with negative-binomial responses. Since the quadratic response model can, even without negative-binomial distribution, simulate overdispersed counts compared to the linear response model, these results were not surprising. In many cases, negative-binomial distributed datasets contained less information than datasets with Poisson-distributed responses, which makes accurate estimation increasingly difficult. The model was not accurate for Bernoulli or ordinal response datasets with small p. Fortunately, data of ecological communities often contain many species. For small n, models fitted to datasets with Bernoulli responses were not accurate, whereas models fit to datasets with ordinal responses showed slightly better performance. This too was not surprising, as datasets with ordinal responses include more information compared to datasets with Bernoulli responses. When the number of sites and species increased above 40, the performance of the GLLVM with quadratic responses in both cases improved considerably. Regardless, especially for Bernoulli responses, the simulated datasets often included too little information for many species to accurately estimate the parameters.

6 | APPLICATIONS TO REAL DATA

We applied the proposed GLLVM with quadratic response model to two different datasets: (a) the well-known hunting spider dataset collected by van der Aart and Smeek-Enserink (1974) in Dutch dunes, available in the `mvabund R` package (Wang et al., 2012), and (b) a dataset of plants in the Swiss Alps (available in the dryad database; D'Amen et al., 2017).

6.1 | Hunting spiders

For the hunting spider dataset, van der Aart and Smeek-Enserink (1974) used pitfall traps to collect spiders over a 60-week period, resulting in a dataset of counts for each of the $n = 28$ sites and $p = 12$ species. It has been used in the testing of ordination methods before (e.g. ter Braak, 1985, 1986; Hui et al., 2015; Yee, 2004), providing the possibility for comparison here. We used the Akaike information criterion corrected for small sample sizes (AICc; Burnham and Anderson, 2002) to find the model that best fitted the hunting spider dataset. We fitted GLLVMs with $d = 1$–3 latent variables, with linear and quadratic responses, including equal, common or unequal tolerances, and fixed row intercepts, all with Poisson or negative-binomial distributions (see Appendix S5 for the details). After selecting the model structure and number of latent variables, we continued to explore different sets of initial values to find the model that maximizes the VA log-likelihood. The best model included $d = 3$ latent variables and unequal tolerances, though a model with unequal tolerances $d = 2$ latent variables and fixed row intercepts was a close second contender (difference of 2.2 in AICc; see Appendix S5). The results for the two latent variables of the final model fit, which explained most residual variation, are presented in Figure 2.

We used the residual variance to determine which latent variables explained most variation, that is, were most important to consider for inference. For the GLLVM with quadratic response model, the first and third latent variables explained most variation in the model; 31% and 58%, respectively, so we will discuss the results of these below. Overall, the GLLVM with quadratic responses explained two and a half times more residual variation than a GLLVM with linear responses and the same number of latent variables. The lengths of the ecological gradients were 5.48 (3.96–7.00, 95% confidence interval), 3.68 (2.65–4.71) and 4.77 (3.10–6.44).

ter Braak (1985) interpreted the first ordination axis of DCA as 'a composite gradient of soil moisture and openness of habitat', as determined by regressing the ordination axis on variables measuring the amount of bare sand, soil moisture and the percentage cover by mosses at sites. Yee (2004) concluded that reflection of the soil surface had the strongest relationship with the first latent variable estimated using a Vector Generalized Additive Model. Similarly, the first latent variable in the GLLVM here has a strong relation with reflection of the surface (correlation coefficient of 0.83), the percentage cover of moss (0.82) and the cover of fallen leaves (−0.75). The second latent variable was related to the cover provided by the herb layer (0.70), and the third latent variable with soil water content (0.77).

ter Braak (1985) and Yee (2004) both visualized quadratic curves of the first latent variable using variations of Poisson regression and Generalized Additive Models respectively. There are clear similarities between the height and the location of species response curves for the first latent variable, and the corresponding response curves described by ter Braak (1985) and Yee (2004). Similarly, Figure 2 here shows a similar arrangement of species as Figure 1 in ter Braak (1986).

ter Braak (1985) concluded that most species exhibited unimodal curves on the first latent variable, though the benefit of a quadratic response model was least to the species *Alopecosa fabrilis*, *Arctosa perita* and *Pardosa lugubris*. Similarly, the optimum of *Pardosa lugubris* could not be estimated by VGAM. Here, as in Yee (2004), *Pardosa lugubris* and *Trochosa terricola* were the most abundant species. On the first latent variable, only the optima of *Pardosa lugubris* and *Pardosa monticola* were located outside the range of the latent variable. On the third latent variable, only the optima of *Arctosa lutetiana* were unobserved. Similar to the conclusion by ter Braak (1985), the confidence intervals for the quadratic coefficients

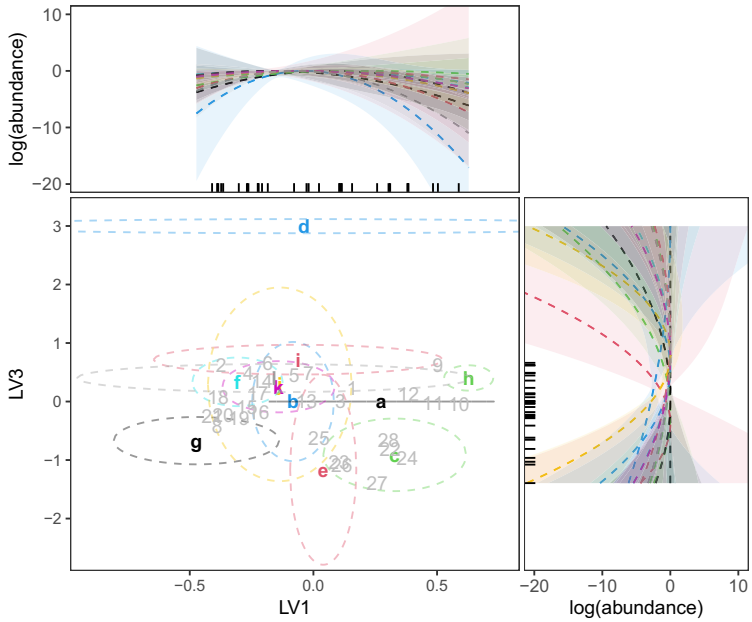


FIGURE 2 Ordination plot for the first two latent variables of the final GLLVM fit to the hunting spider dataset, scaled by the residual variances. Species optima are shown as letters, indicating the following species: a = *Alopecosa accentuata*, b = *Alopecosa cuneata*, c = *Alopecosa fabrilis*, d = *Arctosa lutetiana*, e = *Arctosa perita*, f = *Alonia albimana*, g = *Pardosa lugubris*, h = *Pardosa monticola*, i = *Pardosa nigriceps*, j = *Pardosa pullata*, k = *Trochosa terricola*, l = *Zora spinimana*. Ellipses represent the precision of the ecological niche, which can be interpreted as 'narrowness', so that large or wide ellipses represent species with narrow response curves. Species quadratic curves are included as side panels, with 95% confidence interval bands. Site locations are represented by grey numbers, though prediction regions have not been included in favour of readability

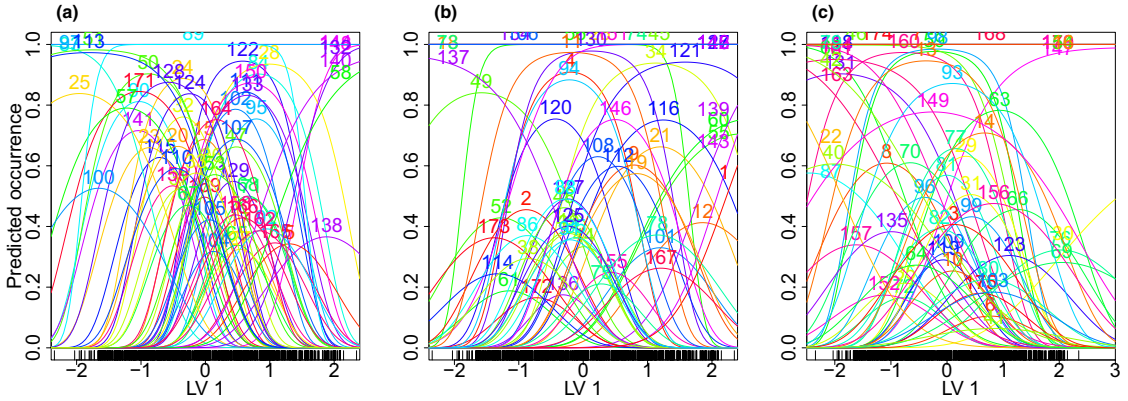


FIGURE 3 One-dimensional figures for the GLLVM fit to the Swiss alpine plants dataset. Each plot includes approximately one third of the species in the dataset, which have been sorted based on their variation explained, so that the first plot includes species explaining most of the variation. Plot (a) represents 63% of the residual variation, plot (b) represents 27% of the residual variation, and plot (c) represents 10% of the residual variation for the first latent variable. The rug plot at the bottom indicates predicted locations of the sites. The numbers correspond with the species names in Figure 4

of *Pardosa lugubris* and *Arctosa perita* included zero on all latent variables, in addition to *Arctosa lutetiana*. From all species on all latent variables, *Arctosa lutetiana* had the smallest tolerance (0.33, on the first latent variable).

6.2 | Swiss alpine plants

In the second application, n = 912 plots of 4 m^2 each were used to record binary data on p = 175 plant species. More species were recorded, but in the original study of this dataset species with less than 22 presences were excluded (D'Amen et al., 2018). Though fitting the model with these species would not have presented any computational issues, their estimates could not necessarily be expected to be accurate. Plots were located on a strong elevation gradient ranging from 375 m to 3,210 m a.s.l. (D'Amen et al., 2018). To improve computation time, we excluded 72 plots without any presences, and 103 plots with less than six presences, so that the final dataset included n = 737 plots.

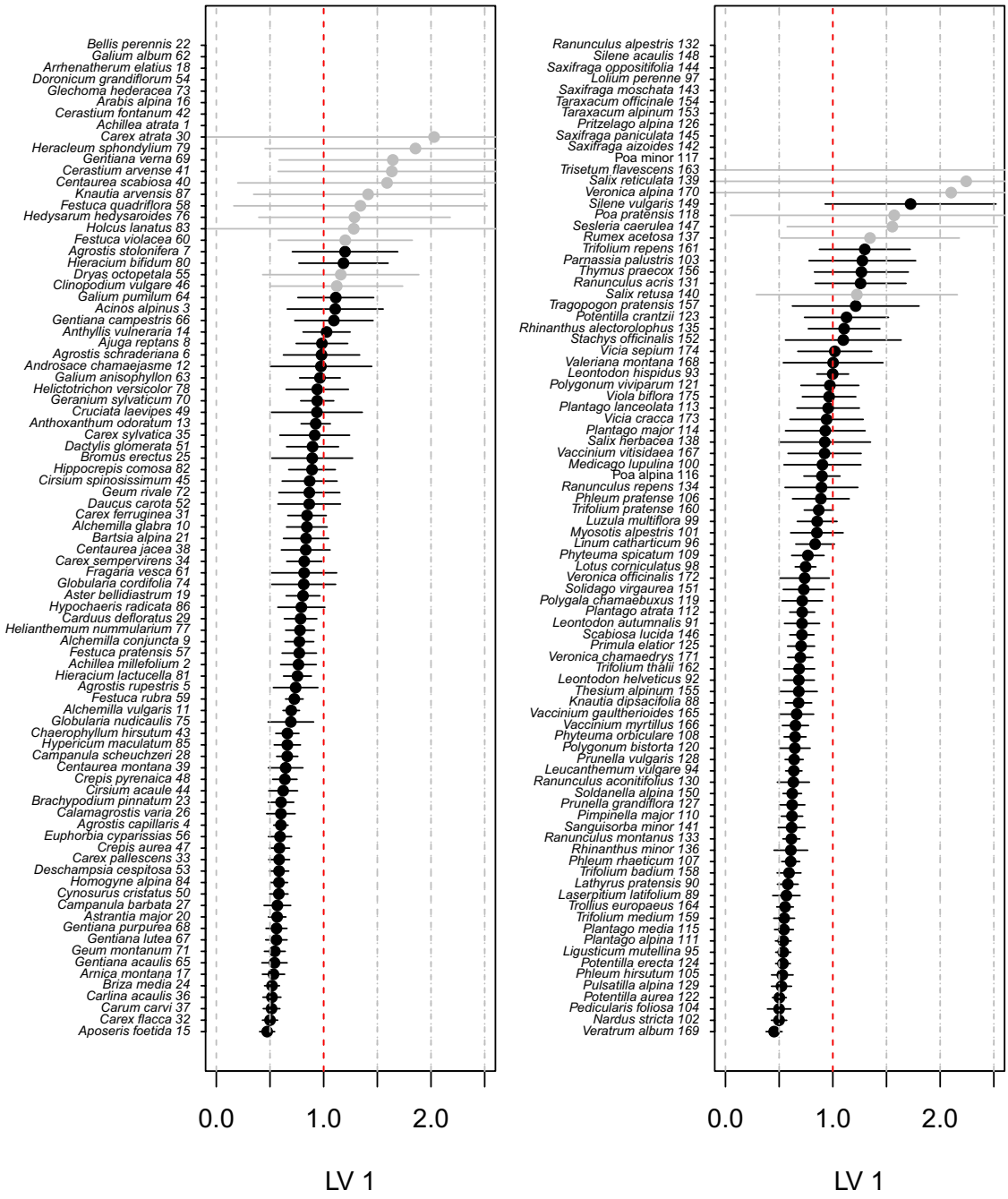


FIGURE 4 Species tolerances and approximate 95% confidence intervals derived using the Delta method, of the first latent variable from the model with unequal tolerances, fitted to the Swiss plants dataset. When tolerances cross 1 (indicated with a red dashed line), species have partially unobserved niches (regardless of the location of their optima). The panels show the first and second half of species in the dataset, respectively, ordered by the size of their tolerances. Species of which the confidence interval for the quadratic coefficients crosses 0 are shown in grey. Species at the top of the plot, seemingly without tolerances, exhibit near linear responses, so that their tolerances are very large. Grey dashed lines are added at increments of 0.5 as visual aid

Instead of selecting the optimal number of latent variables, we directly fitted the proposed GLLVM with quadratic response model to the data, using a Bernoulli distribution and with $d = 2$ latent variables, for the purpose of constructing an ordination diagram. We tested different sets of initial values and retained the model that had the highest log-likelihood.

The first latent variable explained 75% of the overall residual variation in the model, of which 50% was accounted for by the linear term. The length of the first ecological gradient was 4.79 (3.94–5.64, 95% confidence interval), and the length of the second ecological gradient 3.66 (2.86–4.45). Since the first latent variable explained considerably more residual variation than the second, we here focus our inference on that alone for illustration purposes. The species response curves for the first latent variable are visualized in Figure 3a–c. To improve readability, species are numbered by their location in the dataset, for which the corresponding names are included in Figure 4, which also shows species tolerances for the first latent variable, with approximate 95% confidence intervals.

The original dataset additionally included multiple predictor variables, measuring the growing degree-days above zero, a moisture index, total solar radiation over the year, slope, topography and elevation (van der Veen et al., 2021). In an attempt to identify the ecological gradient represented by the first latent variable, we post hoc calculated correlation coefficients between the predictors and the first latent variable. From all predictor variables, elevation was most correlated with the first latent variable (a correlation coefficient of 0.93), though this was collinear with growing degree-days above zero and the moisture index. We additionally fitted two unconstrained GLLVMs with linear species responses and with two latent variables, one of which included a random row intercept, and again calculated a correlation coefficient between the latent variables and elevation. Jamil and ter Braak (2013) showed that a mixed-effects model with random row intercept can account for the squared term of the latent variable. Here, the random row intercept was indeed related to the square of the first latent variable (correlation coefficient of −0.82). The GLLVM with linear species responses but without a row intercept estimated the ecological gradient less successfully (highest correlation coefficient with the elevation predictor of −0.71), than when a row intercept was included (highest correlation coefficient of 0.92). To test more explicitly for the effect of elevation, we additionally fitted a GLLVM with quadratic latent responses and elevation included as a predictor (both the linear and quadratic term, but without sign constraints, though most species exhibited concave curves), and with two latent variables. Including the predictor variable reduced the residual variance to 36% of that in the unconstrained model. The results presented here are from the unconstrained model, though the effect of elevation is presented in Appendix S5, Figure S1.

Of the $p = 175$ species included in the model, 36 had optima that were unobserved, of which 20 were larger than 10 or smaller than −10. The environmental tolerances from species of which the confidence interval for the quadratic coefficients on the first latent variable did not include zero ranged from 0.45 (*Veratrum album*) to 1.72 (*Silene vulgaris*) with a median tolerance of 0.73 and a standard deviation of 0.22. We examined groups of plants at the extremes of the gradient, that is, plants that had optima of minus two or smaller, and plants with optima of two or larger, to further investigate whether the estimated latent variable from the GLLVM with quadratic responses represented an elevation gradient. This approach allowed us to distinguish two groups of plants, the first indicative of lowlands (see Figure 3). In contrast, plant species included on the opposite side of the latent variable were clearly indicative of alpine conditions. Here, we focus our inference on the alpine plants, as those are likely to be most affected by climate change (Walther et al., 2005). All species with optima larger than 2 had confidence intervals for the quadratic coefficients that included zero. Three alpine species had optima located between 1.5 and 2: *Androsace chamaejasme* (1.85, −0.10 to 3.81), *Polygonum viviparum* (1.59, 0.77–2.40) and *Salix herbacea* (1.88, −0.04 to 3.80). Of these three species, *Salix herbacea* had the lowest maximum: −0.34. All three species had a wide response curve on the first latent variable, with tolerances near 1.

Figure 4 clearly shows some species that have smaller tolerances, thus more specialized species are present in the dataset. Six species had a tolerance of 0.50 or smaller: *Aposeris foetida*, *Carex flacca*, *Nardus stricta*, *Pedicularis foliosa*, *Potentilla aurea* and *Veratrum album*.

7 | DISCUSSION

In this article, we extended the GLLVM approach of Hui et al. (2015), to estimate the niches of species with quadratic responses to unobserved ecological gradients. We fitted and performed inference for the GLLVM with quadratic response model by extending the VA approach from Hui et al. (2017). The relation between latent variable models (i.e. unobserved ecological gradients) and ecological niches has been well-described for classical ordination methods (ter Braak & Prentice, 1988; Jongman et al., 1995), yet a method (either classical or model-based) to perform unconstrained (residual) ordination without limiting assumptions for species tolerances has not been available to date.

The similarity in responses of species to unobserved environments can be assessed by examining optima and tolerances, for example, visually using an ordination diagrams, to identify overlap in species distributions, or alternatively by examining a matrix of residual correlations between species. Determining if species exhibit fully quadratic curves in response to ecological gradients, whether tolerances are the same for all species per ecological gradient, or if the equal tolerances assumption is suited for a dataset, comes down to a problem of model selection for GLLVMs. To that end, future research can further investigate approaches such as regularization (e.g. possibly extending the approach of Hui et al., 2018), hypothesis testing or the use of confidence intervals of the quadratic coefficients. Similar to DCA, the GLLVM with quadratic response model provides estimates of gradient length. Here, gradient length is calculated from the quadratic coefficients, which are estimated via a variational approximation approach to maximizing the marginal likelihood function.

For datasets with 50 species and 50 sites or more, the GLLVM with separate quadratic responses for all species accurately retrieved ecological gradients and species-specific parameters, though for continuous responses or counts it was possible to accurately estimate parameters with fewer species or sites. In general, when fitting the GLLVM with quadratic response model to binary or ordinal responses, more information is required than for other data types (similarly as reported in Yee, 2004). However, this is conditional on the information content in a dataset, and the number of required sites and species here should only be considered as a rough rule of thumb. For observed environmental variables, ter Braak and Looman (1986) reported from simulations on estimates of species optima by weighted averaging that, 'with 10–13 presences, the variances of species optima are appreciable'. In our simulations, even with the number of sites fixed at $n = 100$, 24% of species had 13 or fewer presences, indicating difficulty in achieving a sufficient information content in presence–absence datasets to accurately estimate species optima.

We studied the response curves of species to ecological gradients for hunting spiders in a Dutch dune ecosystem (van der Aart & Smeek-Enserink, 1974), and for Swiss alpine plants (D'Amen et al., 2017), using the GLLVM with quadratic response model. Various specialist species can be identified in both datasets, as species with small tolerances on one or multiple latent variables. Specialist species are more likely to be affected by future changes in the environment, and as such their identification is of critical importance to community ecology, to better focus recommendations for conservation efforts.

Modelling rare species is often difficult in community ecology as few ordination methods have the capability to explicitly do so. The quadratic response model has great potential for community ecology, as it can simultaneously accommodate common (large tolerances and maximum i.e. a wide and high niche) and rare species (small tolerances and maximum i.e. a narrow and low niche). The quadratic response model naturally predicts species with unobserved optima, narrow niches and small maxima to have the fewest observations. Since the quadratic response model includes two species-specific parameters per latent variable, and thus requires more information in the data for accurate estimation of parameters than when assuming linear species responses, it potentially requires a large dataset to include sufficient information on rare species and accurately estimate the corresponding parameters. However, the example in this paper using the dataset of counts for hunting spiders (van der Aart & Smeek-Enserink, 1974) suggests that a GLLVM with quadratic response model can be feasible to fit even to small datasets. Regardless, assuming quadratic coefficients to be the same for all species per latent variable might be more realistic for many ecological datasets, while still providing the benefit of an explicit quadratic response model, with all the benefits it provides—calculating species optima, tolerances, maxima, gradient length and their corresponding statistical uncertainties. An additional advantage of a GLLVM-type approach is the ability to use information from both common and rare species to improve estimation of ecological gradients. Even if optima of species with too few observations cannot be accurately estimated, species preferences can be identified based on the ecological gradient, in relation to the response curve of more common species, and based on the direction of the maximum (slope). Without penalization or borrowing information for estimation from more abundant species though, the (quadratic) coefficients for species with few observations are not necessarily expected to be accurate.

An easy-to-use implementation of the quadratic response model with GLLVMs is available in the `gllvm` R package (Niku et al., 2020).

ACKNOWLEDGEMENTS

The authors thank the Spatial Ecology Group at the University of Lausanne, and Manuela D'Amen in specific, for providing the elevation data from the Swiss Alpine plants dataset. The elevation data were originally retrieved from the Swiss Federal Office of Topography. They thank Cajo ter Braak and an anonymous reviewer for helpful comments on earlier drafts of the manuscript. B.V. was supported by a scholarship from the Research Council of Norway (grant number 272408/F40). F.K.C.H. was supported by two Australian Research Council Discovery grants.

AUTHORS' CONTRIBUTIONS

B.v.d.V., K.A.H. and R.B.O. conceived the ideas; B.v.d.V., F.K.C.H. and R.B.O. designed the methodology. All the authors contributed to the writing, reviewing and editing of the draft and gave final approval for publication.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13595.

DATA AVAILABILITY STATEMENT

The hunting spider dataset from the first example is available in the mvabund R package (Wang et al., 2012). The Swiss alpine plants dataset from the second example is available for download in the dryad database (D'Amen et al., 2017), with separate elevation data (van der Veen et al., 2021).

ORCID

Bert van der Veen https://orcid.org/0000-0003-2263-3880
Francis K. C. Hui https://orcid.org/0000-0003-0765-3533
Knut A. Hovstad https://orcid.org/0000-0002-7108-0787
Erik B. Solbu https://orcid.org/0000-0002-6023-3100
Robert B. O'Hara https://orcid.org/0000-0001-9737-3724

REFERENCES

Björk, J. R., Hui, F. K. C., O'Hara, R. B., & Montoya, J. M. (2018). Uncovering the drivers of host-associated microbiota with joint species distribution modelling. *Molecular Ecology*, *27*, 2714–2724. https://doi.org/10.1111/mec.14718

Blanchet, F. G., Cazelles, K., & Gravel, D. (2020). Co-occurrence is not evidence of ecological interactions. *Ecology Letters*, *23*, 1050–1063. https://doi.org/10.1111/ele.13525

Borcard, D., Legendre, P., & Drapeau, P. (1992). Partialling out the spatial component of ecological variation. *Ecology*, *73*, 1045–1055. https://doi.org/10.2307/1940179

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). Springer-Verlag.

D'Amen, M., Mod, H. K., Gotelli, N. J., & Guisan, A. (2017). Disentangling biotic interactions, environmental filters, and dispersal limitation as drivers of species co-occurrence. *Dryad*, https://doi.org/10.5061/dryad.8mv11

D'Amen, M., Mod, H. K., Gotelli, N. J., & Guisan, A. (2018). Disentangling biotic interactions, environmental filters, and dispersal limitation as drivers of species co-occurrence. *Ecography*, *41*, 1233–1244. https://doi.org/10.1111/ecog.03148

Damgaard, C., Hansen, R. R., & Hui, F. K. C. (2020). Model-based ordination of pin-point cover data: Effect of management on dry heathland. *bioRxiv*, 2020.03.05.980060.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, *58*, 453–467. https://doi.org/10.1093/biomet/58.3.453

Gauch, H. G. (1982). *Multivariate analysis in community ecology*. Cambridge University Press.

Halvorsen, R. (2012). A gradient analytic perspective on distribution modelling. *Sommerfeltia*, *35*, 1–165. https://doi.org/10.2478/v10208-011-0015-3

Hill, M. O., & Gauch, H. G. (1980). Detrended correspondence analysis: An improved ordination technique. *Vegetatio*, *42*, 47–58.

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

Hui, F. K. C., Tanaka, E., & Warton, D. I. (2018). Order selection and sparsity in latent variable models via the ordered factor LASSO. *Biometrics*, *74*, 1311–1319. https://doi.org/10.1111/biom.12888

Hui, F. K. C., Taskinen, S., Pledger, S., Foster, S. D., & Warton, D. I. (2015). Model-based approaches to unconstrained ordination. *Methods in Ecology and Evolution*, *6*, 399–411. https://doi.org/10.1111/2041-210X.12236

Hui, F. K. C., Warton, D. I., Ormerod, J. T., Haapaniemi, V., & Taskinen, S. (2017). Variational approximations for generalized linear latent variable models. *Journal of Computational and Graphical Statistics*, *26*, 35–43. https://doi.org/10.1080/10618600.2016.1164708

Inoue, K., Stoeckl, K., & Geist, J. (2017). Joint species models reveal the effects of environment on community assemblage of freshwater mussels and fishes in European rivers. *Diversity and Distributions*, *23*, 284–296. https://doi.org/10.1111/ddi.12520

Jamil, T., & ter Braak, C. J. F. (2013). Generalized linear mixed models can detect unimodal species-environment relationships. *PeerJ*, *1*, e95. https://doi.org/10.7717/peerj.95

Jongman, R., ter Braak, C., & van Tongeren, O. (Eds.) (1995). *Data analysis in community and landscape ecology*. Cambridge University Press.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., & Bell, B. (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, *70*, 1–21.

Lacoste, É., Weise, A. M., Lavoie, M.-F., Archambault, P., & McKindsey, C. W. (2019). Changes in infaunal assemblage structure influence nutrient fluxes in sediment enriched by mussel biodeposition. *Science of the Total Environment*, *692*, 39–48. https://doi.org/10.1016/j.scitotenv.2019.07.235

MacArthur, R., & Levins, R. (1967). The limiting similarity, convergence, and divergence of coexisting species. *The American Naturalist*, *101*, 377–385. https://doi.org/10.1086/282505

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*, 133–142.

Niku, J., Brooks, W., Herliansyah, R., Hui, F. K. C., Taskinen, S., & Warton, D. I. (2019). Efficient estimation of generalized linear latent variable models. *PLoS One*, *14*, e0216129. https://doi.org/10.1371/journal.pone.0216129

Niku, J., Brooks, W., Herliansyah, R., Hui, F. K. C., Taskinen, S., Warton, D. I., van der Veen, B. (2020). Gllvm: Generalized linear latent variable models. https://github.com/JenniNiku/gllvm

Oksanen, J., & Tonteri, T. (1995). Rate of compositional turnover along gradients and total gradient length. *Journal of Vegetation Science*, *6*, 815–824. https://doi.org/10.2307/3236395

Øland, R. H. (1999). On the variation explained by ordination and constrained ordination axes. *Journal of Vegetation Science*, *10*, 131–136.

Øland, R. H., & Eilertsen, O. (1994). Canonical Correspondence Analysis with variation partitioning: Some comments and an application. *Journal of Vegetation Science*, *5*, 117–126.

Ormerod, J. T., & Wand, M. P. (2010). Explaining variational approximations. *The American Statistician*, *64*, 140–153. https://doi.org/10.1198/tast.2010.09058

Ormerod, J. T., & Wand, M. P. (2012). Gaussian variational approximate inference for generalized linear mixed models. *Journal of Computational and Graphical Statistics*, *21*, 2–17. https://doi.org/10.1198/jcgs.2011.09118

Ovaskainen, O., Tikhonov, G., Norberg, A., Blanchet, F. G., Duan, L., Dunson, D., Roslin, T., & Abrego, N. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters*, *20*, 561–576. https://doi.org/10.1111/ele.12757

Paul, W. (2020). Covariate-adjusted species response curves derived from long-term macroinvertebrate monitoring data using classical and contemporary model-based ordination methods. *Ecological Informatics*, *60*, 101159. https://doi.org/10.1016/j.ecoinf.2020.101159

Peres-Neto, P. R., & Jackson, D. A. (2001). How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. *Oecologia*, *129*, 169–178. https://doi.org/10.1007/s004420100720

Pichler, M., & Hartig, F. (2020). A new method for faster and more accurate inference of species associations from novel community data. arXiv:2003.05331 [q-bio, stat]. Retrieved from http://arxiv.org/abs/2003.05331

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

ter Braak, C. J. F. (1985). Correspondence analysis of incidence and abundance data: Properties in terms of a unimodal response model. *Biometrics*, *41*, 859–873. https://doi.org/10.2307/2530959

ter Braak, C. J. (1986). Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis. *Ecology*, *67*, 1167–1179. https://doi.org/10.2307/1938672

ter Braak, C. J. F., & Looman, C. W. N. (1986). Weighted averaging, logistic regression and the Gaussian response model. *Vegetatio*, *65*, 3–11. https://doi.org/10.1007/BF00032121

ter Braak, C. J. F., & Prentice, I. C. (1988). In M. Begon, A. H. Fitter, E. D. Ford, & A. Macfadyen (Eds.), *A theory of gradient analysis*. *Advances in ecological research* (pp. 271–317). Academic Press.

Tikhonov, G., Ovaskainen, O., Oksanen, J., de Jonge, M., Opedal, O., & Dallas, T. (2021). Hmsc: Hierarchical model of species communities. https://CRAN.R-project.org/package=Hmsc

Tobler, M. W., Kéry, M., Hui, F. K. C., Guillera-Arroita, G., Knaus, P., & Sattler, T. (2019). Joint species distribution models with species correlations and imperfect detection. *Ecology*, *100*, e02754. https://doi.org/10.1002/ecy.2754

van der Aart, P., & Smeek-Enserink, N. (1974). Correlations between distributions of hunting spiders (Lycosidae, Ctenidae) and environmental

characteristics in a dune area. *Netherlands Journal of Zoology*, *25*, 1–45.
van der Veen, B., Hui, F. K. C., Hovstad, K. A., Solbu, E. B., & O'Hara, R. B. (2021). Data from: {Model}-based ordination for species with unequal niche widths. *Dryad*, https://doi.org/10.5061/dryad.pnvx0k6m1
Walker, S. C., & Jackson, D. A. (2011). Random-effects ordination: Describing and predicting multivariate correlations and co-occurrences. *Ecological Monographs*, *81*, 635–663. https://doi.org/10.1890/11-0886.1
Walther, G.-R., Beißner, S., & Burga, C. A. (2005). Trends in the upward shift of alpine plants. *Journal of Vegetation Science*, *16*, 541–548. https://doi.org/10.1111/j.1654-1103.2005.tb02394.x
Wang, Y., Naumann, U., Wright, S. T., & Warton, D. I. (2012). Mvabund an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution*, *3*, 471–474.
Warton, D. I., Blanchet, F. G., O'Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C., & Hui, F. K. C. (2015). So many variables: Joint modeling in community ecology. *Trends in Ecology & Evolution*, *30*, 766–779. https://doi.org/10.1016/j.tree.2015.09.007
Wehrden, H. V., Hanspach, J., Bruelheide, H., & Wesche, K. (2009). Pluralism and diversity: Trends in the use and application of ordination methods 1990–2007. *Journal of Vegetation Science*, *20*, 695–705. https://doi.org/10.1111/j.1654-1103.2009.01063.x
Yee, T. W. (2004). A new technique for maximum-likelihood canonical Gaussian ordination. *Ecological Monographs*, *74*, 685–701. https://doi.org/10.1890/03-0078
Zurell, D., Pollock, L. J., & Thuiller, W. (2018). Do joint species distribution models reliably detect interspecific interactions from co-occurrence data in homogenous environments? *Ecography*, *41*, 1812–1819. https://doi.org/10.1111/ecog.03315
Zurell, D., Zimmermann, N. E., Gross, H., Baltensweiler, A., Sattler, T., & Wüest, R. O. (2020). Testing species assemblage predictions from stacked and joint species distribution models. *Journal of Biogeography*, *47*, 101–113. https://doi.org/10.1111/jbi.13608

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: van der Veen B, Hui FKC, Hovstad KA, Solbu EB, O'Hara RB. Model-based ordination for species with unequal niche widths. *Methods Ecol Evol.* 2021;12:1288–1300. https://doi.org/10.1111/2041-210X.13595

Model-based ordination for species with unequal niche widths

Bert van der Veen[123] Francis K.C. Hui[4] Knut A. Hovstad[35] Erik B. Solbu [1]
Robert B. O'Hara[23]

[1]Department of Landscape and Biodiversity, Norwegian Institute of Bioeconomy research, Trondheim, Norway
[2]Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway
[3]Department of Biology, Centre of Biodiversity Dynamics, Norwegian University of Science and Technology, Trondheim, Norway
[4]Research School of Finance, Actuarial Studies and Statistics, Australian National University, Canberra, Australia
[5]The Norwegian Biodiversity Information Centre, Trondheim, Norway

Appendix S1: Residual covariance

Here, we provide a derivation for the residual covariance of a GLLVM with quadratic response model. For a vector of latent variables $\boldsymbol{z}_i$, which is assumed to follow a multivariate standard normal distribution i.e. $\boldsymbol{z}_i \sim \mathcal{N}(\mathbf{0}, \mathrm{I})$, where sites are assumed to be independent, and for the linear predictor $\eta_{ij} = C_{ij} + \boldsymbol{z}_i^\top \boldsymbol{\gamma}_j - \boldsymbol{z}_i^\top \boldsymbol{D}_j \boldsymbol{z}_i$, where C_{ij} is a general quantity that is constant with respect to the latent variables, with $\boldsymbol{\gamma}_j$ a vector of species coefficients for the linear term of the $q = 1 \dots d$ latent variables, and $\boldsymbol{D}_j$ a species-specific positive-definite diagonal matrix of size $d \times d$ including a species coefficients for the quadratic term of the latent variables, the entries of the residual covariance matrix $\boldsymbol{\Sigma}$ for species $j, l = 1 \dots p$, are given by:

$$\begin{aligned}\operatorname{cov}(\boldsymbol{z}_i^\top\boldsymbol{\gamma}_j - \boldsymbol{z}_i^\top\boldsymbol{D}_j\boldsymbol{z}_i, \boldsymbol{z}_i^\top\boldsymbol{\gamma}_l - \boldsymbol{z}_i^\top\boldsymbol{D}_l\boldsymbol{z}_i) &= \operatorname{cov}(\boldsymbol{z}_i^\top\boldsymbol{\gamma}_j, \boldsymbol{z}_i^\top\boldsymbol{\gamma}_l) \\ &+ \operatorname{cov}(\boldsymbol{z}_i^\top\boldsymbol{\gamma}_j, -\boldsymbol{z}_i^\top\boldsymbol{D}_l\boldsymbol{z}_i) \\ &+ \operatorname{cov}(\boldsymbol{z}_i^\top\boldsymbol{\gamma}_l, -\boldsymbol{z}_i^\top\boldsymbol{D}_j\boldsymbol{z}_i) \\ &+ \operatorname{cov}(-\boldsymbol{z}_i^\top\boldsymbol{D}_j\boldsymbol{z}_i, -\boldsymbol{z}_i^\top\boldsymbol{D}_l\boldsymbol{z}_i).\end{aligned}$$

Since the third order central moments of the multivariate normal distribution are zero, we only have to calculate the first and last term,

$$\begin{aligned}\operatorname{cov}(\boldsymbol{z}_i^\top\boldsymbol{\gamma}_j - \boldsymbol{z}_i^\top\boldsymbol{D}_j\boldsymbol{z}_i, \boldsymbol{z}_i^\top\boldsymbol{\gamma}_l - \boldsymbol{z}_i^\top\boldsymbol{D}_l\boldsymbol{z}_i) &= \operatorname{cov}(\boldsymbol{z}_i^\top\boldsymbol{\gamma}_j, \boldsymbol{z}_i^\top\boldsymbol{\gamma}_l) \\ &+ \operatorname{cov}(-\boldsymbol{z}_i^\top\boldsymbol{D}_j\boldsymbol{z}_i, -\boldsymbol{z}_i^\top\boldsymbol{D}_l\boldsymbol{z}_i) \\ &= \boldsymbol{\gamma}_j^\top\boldsymbol{\gamma}_l + 2\operatorname{tr}(\boldsymbol{D}_j\boldsymbol{D}_l). \qquad (1)\end{aligned}$$

Appendix S2: Variational approximations

In the derivation of the VA log-likelihoods below, we let C_{ij} generically denote a quantity that is constant with respect to the latent variables $\boldsymbol{z}_i$, for example species-specific intercepts β_{0j}, for $i = 1 \dots n$ sites and $j = 1 \dots p$ species. We start by defining the linear predictor:

$$\eta_{ij} = C_{ij} + \boldsymbol{z}_i^\top \boldsymbol{\gamma}_j - \boldsymbol{z}_i^\top \boldsymbol{D}_j \boldsymbol{z}_i,$$

where $\boldsymbol{\gamma}_j$ is a vector of species coefficients for the linear term of the $q = 1 \dots d$ latent variables, and $\boldsymbol{D}_j$ is a species-specific positive-definite diagonal matrix of size $d \times d$. All parameters are collected in $\boldsymbol{\Theta}$, including any nuisance parameters where applicable, such as cutoffs $\boldsymbol{\zeta}_{jk}$ for ordinal responses, and dispersion parameters ϕ_j for the Gaussian, gamma, and negative-binomial responses, i.e. $\boldsymbol{\Theta} = (C_{11} \dots C_{ij}, \gamma_{11} \dots \gamma_{jq}, D_{111} \dots D_{jqq}, \phi_1 \dots \phi_j)^\top$.

Variational approximations

The GLLVM log-likelihood is:

$$\mathcal{L}(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \log \Bigg\{ \int \prod_{j=1}^{p} f\Big(y_{ij} | \boldsymbol{z}_i, \boldsymbol{\Theta}\Big) h\Big(\boldsymbol{z}_i\Big) d\boldsymbol{z}_i \Bigg\}, \tag{2}$$

where $\boldsymbol{z}_i \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I})$, and where $f(y_{ij} | \boldsymbol{z}_i, \boldsymbol{\Theta})$ is a GLM-type distribution (e.g. in the exponential family). Variational approximations (VA) applies Jensen's inequality to construct a lower bound for the marginal log-likelihood [see; Hui et al. (2017); Ormerod and Wand (2010)] giving:

$$\log \mathcal{L}_{VA}(\boldsymbol{\Theta}, \boldsymbol{\xi}) = \sum_{i=1}^{n} \sum_{j=1}^{p} \Bigg\{ \int \log \Bigg(\frac{f(y_{ij} | \boldsymbol{z}_i, \boldsymbol{\Theta}) h(\boldsymbol{z}_i)}{\mathrm{q}(\boldsymbol{z}_i | \boldsymbol{a}_i, \boldsymbol{A}_i)} \Bigg) \mathrm{q}(\boldsymbol{z}_i | \boldsymbol{a}_i, \boldsymbol{A}_i) d\boldsymbol{z}_i \Bigg\}, \tag{3}$$

where $\mathrm{q}(\boldsymbol{z}_i | \boldsymbol{a}_i, \boldsymbol{A}_i)$ is a variational distribution of the latent variables, which we assume to be multivariate normal with mean $\boldsymbol{a}_i$ and covariance matrix $\boldsymbol{A}_i$, collected in $\boldsymbol{\xi}$, i.e. $\boldsymbol{\xi} = \{a_{11} \dots a_{iq}, \mathrm{vech}(\boldsymbol{A}_1) \dots \mathrm{vech}(\boldsymbol{A}_i)\}^\top$, where $\mathrm{vech}(\cdot)$ is a half-vectorizing operator, converting a symmetric matrix to a vector, by retrieving its lower triangular entries. The variational distribution serves as a closed form approximation to the true posterior distribution of the latent variables. Note that in general, an optimal variational distribution (in the sense of minimizing the Kullback-Leibler divergence to the true posterior) of the latent variables can be found by following equation (5) from Ormerod and Wand (2010). On the other hand, to ensure that a tractable lower bound is obtained to facilitate efficient computation for the proposed GLLVM with quadratic response model, we can instead choose a simpler parametric form for the variational distribution, and this is the

approach we adopt here by choosing $q(\boldsymbol{z}_i)$ to be a multivariate normal distribution. As an aside, note that as a consequence of the VA framework (regardless of the form for the variational distribution chosen), the Kullback-Leibler divergence between the true posterior distribution of the latent variables and the variational distribution is minimized. This is useful, in that it at least ensures that the variational parameters estimated best represent the true posterior distribution (in a Kullback-Leibler sense; see Ormerod and Wand (2010) or Hui et al. (2017) for more details). The VA log-likelihood is then calculated by taking expectations over the components of equation (3):

$$\log \mathcal{L}_{VA}(\boldsymbol{\Theta}, \boldsymbol{\xi}) = \sum_{i=1}^{n}\sum_{j=1}^{p} \mathrm{E}\left[\frac{y_{ij}\eta_{ij} - \mathrm{b}\{\eta_{ij}\}}{\mathrm{a}\{\phi_j\}} + c\{y_{ij}, \phi_j\}\right] - \frac{1}{2}\sum_{i=1}^{n} \mathrm{E}\left[\boldsymbol{z}_i^\top \boldsymbol{z}_i\right] - \sum_{i=1}^{n} \mathrm{E}\left[\log\left\{\mathrm{q}\left(\boldsymbol{z}_i|\boldsymbol{a}_i, \boldsymbol{A}_i\right)\right\}\right]. \quad (4)$$

For brevity, we summarize the first two terms in equation (4) as $\mathrm{E}\{\mathcal{L}(\boldsymbol{\Theta})\}$. The expectations for the second and third terms in equation (4) follow the standard result:

$$\mathrm{E}[\log\{\mathrm{q}(\boldsymbol{z}_i|\boldsymbol{a}_i, \boldsymbol{A}_i)\}] \propto -\frac{1}{2}\log\det(\boldsymbol{A}_i)$$
$$\mathrm{E}[\boldsymbol{z}_i^\top \boldsymbol{z}_i] = \mathrm{tr}(\boldsymbol{A}_i) + \boldsymbol{a}_i^\top \boldsymbol{a}_i$$

Additionally, taking expectations over the linear predictor from above, with respect to the variational distribution of the latent variables, gives:

$$\tilde{\eta}_{ij} = C_{ij} + \mathrm{E}(\boldsymbol{z}_i^\top)\boldsymbol{\gamma}_j + \mathrm{E}(\boldsymbol{z}_i^\top \boldsymbol{D}_j \boldsymbol{z}_i) = C_{ij} + \boldsymbol{a}_i^\top \boldsymbol{\gamma_j} - \boldsymbol{a}_i^\top \boldsymbol{D}_j \boldsymbol{a}_i - \mathrm{tr}(\boldsymbol{D}_j \boldsymbol{A}_i).$$

These two results above are valid for all distributions, so that the general VA log-likelihood for a GLLVM with quadratic response model is:

$$\log \mathcal{L}_{VA}(\boldsymbol{\Theta}, \boldsymbol{\xi}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\left[\frac{y_{ij}\tilde{\eta}_{ij} - \mathrm{E}\{\mathrm{b}(\eta_{ij})\}}{\mathrm{a}\{\phi_j\}} + c\{y_{ij}, \phi_j\}\right] + \frac{1}{2}\sum_{i=1}^{n}\left\{\log\det\left(\boldsymbol{A}_i\right) - \mathrm{tr}\left(\boldsymbol{A}_i\right) - \boldsymbol{a}_i^\top \boldsymbol{a}_i\right\}, \quad (5)$$

where $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are known functions. Calculating $\mathrm{E}\{\mathrm{b}(\eta_{ij})\}$ is most challenging as the solution has to be derived separately for most distributions. Below follow the derivations for the Gaussian, Poisson, negative-binomial, Bernoulli, ordinal and gamma likelihoods.

Gaussian: continuous responses

The joint log-likelihood for Gaussian distributed responses is:

$$\mathcal{L}(\boldsymbol{\Theta}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\left\{-\frac{1}{2}\log\left(\sigma_j^2\right) - \frac{1}{2\sigma_j^2}\left(y_{ij} - \eta_{ij}\right)^2\right\} - \frac{1}{2}\sum_{i=1}^{n}\boldsymbol{z}_i^\top\boldsymbol{z}_i, \tag{6}$$

where terms constant with respect to the parameters have been omitted. Here, σ_j is a dispersion parameter accounting for the residual variance of the Gaussian distribution (i.e. $\sigma_j = \phi_j$ above). The VA log-likelihood is derived by working our expressions for the expectations in equation (4), with respect to the variational distribution of latent variables:

$$\mathrm{E}[\mathcal{L}\{\boldsymbol{\Theta}\}] = \sum_{i=1}^{n}\sum_{j=1}^{p}\left[-\frac{1}{2}\log\left\{\sigma_j^2\right\} - \frac{1}{2\sigma_j^2}\mathrm{E}\left\{\left(y_{ij} - \eta_{ij}\right)^2\right\}\right] - \frac{1}{2}\sum_{i=1}^{n}\left\{\mathrm{tr}\left(\boldsymbol{A}_i\right) + \boldsymbol{a}_i^\top\boldsymbol{a}_i\right\},$$

$$\mathrm{E}[\log\{\mathrm{q}(\boldsymbol{z}_i|\boldsymbol{a}_i, \boldsymbol{A}_i)\}] = -\frac{1}{2}\log\det(\boldsymbol{A}_i),$$

where terms constant with respect to the parameters have been omitted. The term $\mathrm{E}\{(y_{ij} - \eta_{ij})^2\}$ can instead be written and expanded as:

$$\begin{aligned}\mathrm{E}\{(y_{ij} - \eta_{ij})^2\} &= \mathrm{E}\{(y_{ij} - \tilde{\eta}_{ij} + \tilde{\eta}_{ij} - \eta_{ij})^2\} \\ &= \mathrm{E}\{(y_{ij} - \tilde{\eta}_{ij})^2\} + 2\mathrm{E}\{(y_{ij} - \tilde{\eta}_{ij})(\tilde{\eta}_{ij} - \eta_{ij})\} + \mathrm{E}\{(\tilde{\eta}_{ij} - \eta_{ij})^2\}.\end{aligned}$$

As the first term does not require taking expectations, and since the second term is zero, only the third term has to be calculated, so we note that $\mathrm{E}\{(\tilde{\eta}_{ij} - \eta_{ij})^2\} = \mathrm{var}(\eta_{ij})$, and work out the expectations accordingly:

$$\begin{aligned}\mathrm{var}(\eta_{ij}) &= \mathrm{var}(C_{ij} + \boldsymbol{z}_i^\top\boldsymbol{\gamma}_j - \boldsymbol{z}_i^\top\boldsymbol{D}_j\boldsymbol{z}_i) \\ &= \mathrm{var}(\boldsymbol{z}_i^\top\boldsymbol{\gamma}_j) + \mathrm{var}(\boldsymbol{z}_i^\top\boldsymbol{D}_j\boldsymbol{z}_i) - 2\mathrm{cov}(\boldsymbol{z}_i^\top\boldsymbol{\gamma}_j, \boldsymbol{z}_i^\top\boldsymbol{D}_j\boldsymbol{z}_i) \\ &= \mathrm{tr}(\boldsymbol{\gamma}_j\boldsymbol{\gamma}_j^\top\boldsymbol{A}_i) + 2\mathrm{tr}(\boldsymbol{D}_j\boldsymbol{A}_i\boldsymbol{D}_j\boldsymbol{A}_i) + 4\boldsymbol{a}_i^\top\boldsymbol{D}_j\boldsymbol{A}_i\boldsymbol{D}_j\boldsymbol{a}_i - 2\mathrm{cov}(\boldsymbol{z}_i^\top\boldsymbol{\gamma}_j, \boldsymbol{z}_i^\top\boldsymbol{D}_j\boldsymbol{z}_i).\end{aligned} \tag{8}$$

For the last term:

$$
\begin{aligned}
\operatorname{cov}(\boldsymbol{z}_i^\top \boldsymbol{\gamma}_j, \boldsymbol{z}_i^\top \boldsymbol{D}_j \boldsymbol{z}_i) &= \operatorname{cov}\{\boldsymbol{z}_i^\top \boldsymbol{\gamma}_j - \mathrm{E}(\boldsymbol{z}_i^\top \boldsymbol{\gamma}_j), \boldsymbol{z}_i^\top \boldsymbol{D}_j \boldsymbol{z}_i - \mathrm{E}(\boldsymbol{z}_i^\top \boldsymbol{D}_j \boldsymbol{z}_i)\} \\
&= \operatorname{cov}\{\sum_{k=1}^{d} \gamma_{jk}(z_{ik} - a_{ik}), \sum_{l=1}^{d} D_{jll} z_{il}^2 - D_{jll}\mathrm{E}(z_{il}^2)\} \\
&= \sum_{k,l=1}^{d} \operatorname{cov}\{\gamma_{jk}(z_{ik} - a_{ik}), D_{jll} z_{il}^2 - D_{jll}\mathrm{E}(z_{il}^2)\} \\
&= \sum_{k,l=1}^{d} \mathrm{E}[\gamma_{jk}\{z_{ik} - a_{ik}\}\{D_{jll} z_{il}^2 - D_{jll}\mathrm{E}(z_{il}^2)\}] \\
&= \sum_{k,l=1}^{d} \mathrm{E}\{\gamma_{jk}(z_{ik} - a_{ik}) D_{jll} z_{il}^2\} \\
&= \sum_{k,l=1}^{d} \gamma_{jk} D_{jll} \mathrm{E}\{(z_{ik} - a_{ik})(z_{il} - a_{il} + a_{il})^2\} \\
&= \sum_{k,l=1}^{d} \gamma_{jk} D_{jll} \mathrm{E}\{(z_{ik} - a_{ik})(z_{il} - a_{il})^2 + 2a_{il}(z_{ik} - a_{ik})(z_{il} - a_{il})\} \\
&= \sum_{k,l=1}^{d} 2\gamma_{jk} D_{jll} a_{il} \mathrm{E}\{(z_{ik} - a_{ik})(z_{il} - a_{il})\} \\
&= 2 \sum_{k,l=1}^{d} \gamma_{jk} D_{jll} a_{il} A_{ikl} \\
&= 2\boldsymbol{a}_i^\top \boldsymbol{D}_j \boldsymbol{A}_i \boldsymbol{\gamma}_j.
\end{aligned}
\tag{9}
$$

Thus, the Gaussian VA log-likelihood for the quadratic model is:

$$
\begin{aligned}
\mathcal{L}_{VA}(\boldsymbol{\Theta}, \boldsymbol{\xi}) = \sum_{i=1}^{n} \sum_{j=1}^{p} \Bigg[-\frac{1}{2} \log\left\{\sigma_j^2\right\} - \frac{1}{2\sigma_j^2} \Big\{ y_{ij}^2 + \tilde{\eta}_{ij}^2 - 2 y_{ij} \tilde{\eta}_{ij} \\
+ \operatorname{tr}\left(\boldsymbol{\gamma}_j \boldsymbol{\gamma}_j^\top \boldsymbol{A}_i\right) + 2\operatorname{tr}\left(\boldsymbol{D}_j \boldsymbol{A}_i \boldsymbol{D}_j \boldsymbol{A}_i\right) + 4\boldsymbol{a}_i^\top \boldsymbol{D}_j \boldsymbol{A}_i \boldsymbol{D}_j \boldsymbol{a}_i - 4\boldsymbol{a}_i^\top \boldsymbol{D}_j \boldsymbol{A}_i \boldsymbol{\gamma}_j \Big\} \Bigg] \\
+ \frac{1}{2} \sum_{i=1}^{n} \Big\{ \log\det\left(\boldsymbol{A}_i\right) - \operatorname{tr}\left(\boldsymbol{A}_i\right) - \boldsymbol{a}_i^\top \boldsymbol{a}_i \Big\}.
\end{aligned}
\tag{10}
$$

Bernoulli: presence-absence responses

We model presence-absence data with a probit-link function and a Bernoulli distribution, giving the joint log-likelihood:

$$
\mathcal{L}(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \sum_{j=1}^{p} \Big[y_{ij} \log\Big\{ \mathrm{I}\Big(\nu_{ij} \geq 0\Big) \Big\} + \Big\{ 1 - y_{ij} \Big\} \log\Big\{ \mathrm{I}\Big(\nu_{ij} < 0\Big) \Big\} - \frac{1}{2} \Big\{ \nu_{ij} - \eta_{ij} \Big\}^2 \Big] - \frac{1}{2} \sum_{i=1}^{n} \boldsymbol{z}_i^\top \boldsymbol{z}_i, \tag{11}
$$

where terms constant with respect to the parameters have been omitted. Here, v_{ij} is an auxiliary variable included to aid integration, with indicator function $\mathrm{I}(\cdot)$. As in Hui et al. (2017), we assume $v_{ij} \sim \mathcal{N}(\eta_{ij}, 1)$, and by equation (5) in Ormerod and Wand (2010), we determine the optimal variational distribution for the auxiliary variable by taking expectations over equation (11) with respect to ν_{ij}:

$$\begin{aligned}\log\{\mathrm{q}(\nu_{ij})\} &\propto \mathrm{E}\{\mathcal{L}(\boldsymbol{\Theta})\} \\ &\propto \log\{\mathrm{I}(\nu_{ij} > 0)\} - \frac{1}{2}\{\nu_{ij} - \mathrm{E}(\eta_{ij})\}^2, \qquad y_{ij} = 1 \\ &\propto \log\{\mathrm{I}(\nu_{ij} < 0)\} - \frac{1}{2}\{\nu_{ij} - \mathrm{E}(\eta_{ij})\}^2, \qquad y_{ij} = 0,\end{aligned} \tag{12}$$

where terms constant with respect to the parameters have been omitted. The second term of both lines in equation (12) expands to $\nu_{ij}\mathrm{E}(\eta_{ij}) - \frac{1}{2}\nu_{ij}^2 - \frac{1}{2}\mathrm{E}(\eta_{ij}^2)$, showing that the distribution of ν_{ij} does not depend on η_{ij}^2. As equation (12) is still quadratic after taking expectations, the optimal distribution for the auxiliary variable is truncated normal with location parameter $\mathrm{E}(\eta_{ij}) = \tilde{\eta}_{ij}$ and scale parameter one, as in Hui et al. (2017). The distribution has limits $(0,\infty)$ and $(-\infty,0)$ for $y_{ij} = 1$ and $y_{ij} = 0$ respectively.

Due to the inclusion of the auxiliary variable, the components of the VA log-likelihood are now:

$$\mathcal{L}_{VA}(\boldsymbol{\Theta}, \boldsymbol{\xi}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\mathrm{E}[\mathcal{L}\{\boldsymbol{\Theta}\}] - \mathrm{E}[\log\{\mathrm{q}(\nu_{ij})\}] - \sum_{i=1}^{n}\mathrm{E}[\log\{\mathrm{q}(\boldsymbol{z}_i|\boldsymbol{a}_i, \boldsymbol{A}_i)\}]. \tag{13}$$

Then, taking expectations over each of the components in equation (13),

$$\begin{aligned}\mathrm{E}[\mathcal{L}\{\boldsymbol{\Theta}\}] = \sum_{i=1}^{n}\sum_{j=1}^{p}\mathrm{E}\left[y_{ij}\log\left\{\mathrm{I}\left(\nu_{ij} > 0\right)\right\} + \left\{1 - y_{ij}\right\}\log\left\{\mathrm{I}\left(\nu_{ij} < 0\right)\right\}\right] - \frac{1}{2}\mathrm{E}\left[\left\{\nu_{ij} - \eta_{ij}\right\}^2\right] \\ -\frac{1}{2}\sum_{i=1}^{n}\left\{\mathrm{tr}\left(\boldsymbol{A}_i\right) + \boldsymbol{a}_i^\top\boldsymbol{a}_i\right\}\end{aligned} \tag{14}$$

$$\begin{aligned}\mathrm{E}[\log\{\mathrm{q}(\nu_{ij})\}] = \mathrm{E}\left[y_{ij}\log\left\{\mathrm{I}\left(\nu_{ij} > 0\right)\right\} + \left(1 - y_{ij}\right)\log\left\{\mathrm{I}\left(\nu_{ij} < 0\right)\right\}\right] - \frac{1}{2}\mathrm{E}\left[\left\{\nu_{ij} - \tilde{\eta}_{ij}\right\}^2\right] \\ -y_{ij}\log\left\{\Phi\left(\tilde{\eta}_{ij}\right)\right\} - \left(1 - y_{ij}\right)\log\left\{1 - \Phi\left(\tilde{\eta}_{ij}\right)\right\} \\ \mathrm{E}[\log\{\mathrm{q}(\boldsymbol{z}_i|\boldsymbol{a}_i, \boldsymbol{A}_i)\}] = -\frac{1}{2}\log\det(\boldsymbol{A}_i),\end{aligned}$$

where terms constant with respect to the parameters have been omitted. The term $\mathrm{E}[\{\nu_{ij} - \eta_{ij}\}^2]$ can be rewritten to simplify taking expectations, in the following way:

$$\begin{aligned}\mathrm{E}\{(\nu_{ij}-\eta_{ij})^2\} &= \mathrm{E}\{(\nu_{ij}-\tilde{\eta}_{ij}+\tilde{\eta}_{ij}-\eta_{ij})^2\} \\ &= \mathrm{E}\{(\nu_{ij}-\tilde{\eta}_{ij})^2\} + 2\mathrm{E}\{(\nu_{ij}-\tilde{\eta}_{ij})(\tilde{\eta}_{ij}-\eta_{ij})\} + \mathrm{E}\{(\tilde{\eta}_{ij}-\eta_{ij})^2\}.\end{aligned}$$

The first term cancels with the same in the first line of equation (14), the second term is zero under the assumption of independent variational distributions and because η_{ij} is not a function of ν_{ij}. The solution to the last term is given in equation (9).

Thus, the Bernoulli VA log-likelihood for the quadratic model is:

$$\begin{aligned}\mathcal{L}_{VA}(\boldsymbol{\Theta},\boldsymbol{\xi}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\Bigg[y_{ij}\log\Big\{\Phi\Big(\tilde{\eta}_{ij}\Big)\Big\} + \Big\{1-y_{ij}\Big\}\log\Big\{1-\Phi\Big(\tilde{\eta}_{ij}\Big)\Big\} - \frac{1}{2}\mathrm{tr}\Big\{\boldsymbol{\gamma}_j\boldsymbol{\gamma}_j^\top\boldsymbol{A}_i\Big\} - \mathrm{tr}\Big\{\boldsymbol{D}_j\boldsymbol{A}_i\boldsymbol{D}_j\boldsymbol{A}_i\Big\} \\ -2\boldsymbol{a}_i^\top\boldsymbol{D}_j\boldsymbol{A}_i\boldsymbol{D}_j\boldsymbol{a}_i + 2\boldsymbol{a}_i^\top\boldsymbol{D}_j\boldsymbol{A}_i\boldsymbol{\gamma}_j\Bigg] + \frac{1}{2}\sum_{i=1}^{n}\Big\{\log\det\Big(\boldsymbol{A}_i\Big) - \mathrm{tr}\Big(\boldsymbol{A}_i\Big) - \boldsymbol{a}_i^\top\boldsymbol{a}_i\Big\}.\end{aligned} \tag{15}$$

Ordinal: ordered responses

The ordered response model follows from an extension of the Bernoulli VA log-likelihood for multiple categories. We define a $p \times K$ matrix of cutoffs $\boldsymbol{\zeta}_{jk}$, where which serves to introduce order to the probability of occurrence in each of the $k = 1 \dots K$ categories per species j. The first cutoff of each species is set to zero for reasons of parameter identifiability in the presence of species intercepts β_{0j}, and so that $\zeta_{j0} = -\infty$ and $\zeta_{jK} = \infty$, ensuring that the probability of occurrence in the first category is at most $\frac{1}{2}$, so that this parametrization of the ordination distribution corresponds with that of the Bernoulli distribution, when $K = 2$.

With the assumption of species-specific cutoffs as above, comes the requirement of more than one observation per species and category. Missing categories are not allowed, though this requirement can be relaxed, by noting that the categories per species are arbitrary (as they are indexes only), so that per species they may be renumbered to exclude categories that lack any observations (i.e. missing categories). To improve the practical usefulness of the ordinal model, we here introduce an additional, simpler, parametrization where we allow $\zeta_{jk} = \zeta_k$, i.e. species-common cutoffs, so that $\zeta_0 = -\infty$ and $\zeta_K = \infty$. This relaxes the requirement of at least one observation per category per species to at least one observation per category in the whole dataset, which we deem more realistic for real world data. This has the additional benefit of optimally utilising information in the dataset from both frequently and infrequently occurring species, to more accurately estimate the cutoffs.

In either case, the variational distribution for auxiliary variable ν_{ij} is truncated normal, but with limits

ζ_{k-1}, ζ_k for $y_{ijk} > 0$ in the species-common cutoff case. Additionally, we define y_{ijk} as a K-dimensional array, so that the VA log-likelihood with probit link, for ordinal responses and with the quadratic response model is:

$$\mathcal{L}_{VA}(\boldsymbol{\Theta}, \boldsymbol{\xi}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{K_j}\Bigg[y_{ijk}\log\Bigg\{\Phi\Bigg(\zeta_{jk} - \tilde{\eta}_{ij}\Bigg) - \Phi\Bigg(\zeta_{jk-1} - \tilde{\eta}_{ij}\Bigg)\Bigg\} - \frac{1}{2}\mathrm{tr}\Bigg\{\boldsymbol{\gamma}_j\boldsymbol{\gamma}_j^\top\boldsymbol{A}_i\Bigg\} - \mathrm{tr}\Bigg\{\boldsymbol{D}_j\boldsymbol{A}_i\boldsymbol{D}_j\boldsymbol{A}_i\Bigg\}$$
$$-2\boldsymbol{a}_i^\top\boldsymbol{D}_j\boldsymbol{A}_i\boldsymbol{D}_j\boldsymbol{a}_i + 2\boldsymbol{a}_i^\top\boldsymbol{D}_j\boldsymbol{A}_i\boldsymbol{\gamma}_j\Bigg] + \frac{1}{2}\sum_{i=1}^{n}\Bigg\{\log\det\Bigg(\boldsymbol{A}_i\Bigg) - \mathrm{tr}\Bigg(\boldsymbol{A}_i\Bigg) - \boldsymbol{a}_i^\top\boldsymbol{a}_i\Bigg\}. \tag{16}$$

Poisson: counted responses

The joint log-likelihood for Poisson distributed responses with a log-link function is:

$$\mathcal{L}(\boldsymbol{\Theta}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\Bigg\{y_{ij}\eta_{ij} - \exp\Bigg(\eta_{ij}\Bigg)\Bigg\} - \frac{1}{2}\sum_{i=1}^{n}\boldsymbol{z}_i^\top\boldsymbol{z}_i, \tag{17}$$

where terms constant with respect to the parameters have been omitted. The VA log-likelihood is derived by working out explicit expressions for the expectations in equation (4) with respect to the latent variables:

$$\mathrm{E}[\mathcal{L}\{\boldsymbol{\Theta}\}] = \sum_{i=1}^{n}\sum_{j=1}^{p}\Bigg[y_{ij}\mathrm{E}\Bigg\{\eta_{ij}\Bigg\} - \mathrm{E}\Bigg\{\exp\Bigg(\eta_{ij}\Bigg)\Bigg\}\Bigg] - \frac{1}{2}\sum_{i=1}^{n}\Bigg\{\mathrm{tr}(\boldsymbol{A}_i) + \boldsymbol{a}_i^\top\boldsymbol{a}_i\Bigg\}$$

$$\mathrm{E}[\log\{\mathrm{q}(\boldsymbol{z}_i|\boldsymbol{a}_i, \boldsymbol{A}_i)\}] = -\frac{1}{2}\log\det(\boldsymbol{A}_i),$$

where terms constant with respect to the parameters have been omitted. To facilitate integration of $\mathrm{E}\{\exp(\eta_{ij})\}$, we rewrite the quadratic model to:

$$\eta_{ij} = C_{ij} - \frac{1}{2}\Bigg(-2\boldsymbol{z}_i^\top\boldsymbol{\gamma}_j + 2\boldsymbol{z}_i^\top\boldsymbol{D}_j\boldsymbol{z}_i\Bigg),$$

so that the expression that requires integration is:

$$\int\exp\Bigg[C_{ij} - \frac{1}{2}\Bigg\{2\boldsymbol{z}_i^\top\boldsymbol{D}_j\boldsymbol{z}_i - 2\boldsymbol{z}_i\boldsymbol{\gamma}_j + \Bigg(\boldsymbol{z}_i - \boldsymbol{a}_i\Bigg)^\top\boldsymbol{A}_i^{-1}\Bigg(\boldsymbol{z}_i - \boldsymbol{a}_i\Bigg)\Bigg\}\Bigg]\det\Bigg(\boldsymbol{A}_i\Bigg)^{-\frac{1}{2}}d\boldsymbol{z}_i.$$

Taking out terms that are constant with respect to the latent scores, this expression can be rewritten to:

$$\exp\Bigg(C_{ij} - \frac{1}{2}\boldsymbol{a}_i^\top\boldsymbol{A}_i\boldsymbol{a}_i\Bigg)\int\exp\Bigg\{-\frac{1}{2}\Bigg(\boldsymbol{z}_i - v_{ij}\Bigg)^\top\Bigg(2\boldsymbol{D}_j + \boldsymbol{A}_i^{-1}\Bigg)\Bigg(\boldsymbol{z}_i - v_{ij}\Bigg) + \frac{1}{2}\boldsymbol{v}_{ij}^\top\Bigg(2\boldsymbol{D}_j + \boldsymbol{A}_i^{-1}\Bigg)v_{ij}\Bigg\}d\boldsymbol{z}_i,$$

with vector $\boldsymbol{v}_{ij}$ as:

$$\boldsymbol{z}_i^\top \Big(2\boldsymbol{D}_j + \boldsymbol{A}_i^{-1}\Big)\boldsymbol{v}_{ij} = \boldsymbol{z}_i^\top \boldsymbol{\gamma}_j + \boldsymbol{z}_i^\top \boldsymbol{A}_i^{-1}\boldsymbol{a}_i$$

$$\begin{aligned}\boldsymbol{v}_{ij} &= \frac{1}{\boldsymbol{z}_i^\top}\Big(2\boldsymbol{D}_j + \boldsymbol{A}_i^{-1}\Big)^{-1}\Big(\boldsymbol{z}_i^\top \boldsymbol{\gamma}_j + \boldsymbol{z}_i^\top \boldsymbol{A}_i^{-1}\boldsymbol{a}_i\Big) \\ &= \Big(2\boldsymbol{D}_j + \boldsymbol{A}_i^{-1}\Big)^{-1}\Big(\boldsymbol{\gamma}_j + \boldsymbol{A}_i^{-1}\boldsymbol{a}_i\Big).\end{aligned}$$

Now, since the term $\exp\{\frac{1}{2}\boldsymbol{v}_{ij}^\top(2\boldsymbol{D}_j + \boldsymbol{A}_i^{-1})\boldsymbol{v}_{ij}\}$ is constant with respect to the latent variables too, we can again rewrite the expression:

$$\exp\Big[C_{ij} + \frac{1}{2}\Big\{\boldsymbol{v}_{ij}^\top\Big(2\boldsymbol{D}_j + \boldsymbol{A}_i^{-1}\Big)\boldsymbol{v}_{ij} - \boldsymbol{a}_i^\top \boldsymbol{A}_i \boldsymbol{a}_i\Big\}\Big]\int \exp\Big\{-\frac{1}{2}\Big(\boldsymbol{z}_i - \boldsymbol{v}_{ij}\Big)^\top\Big(2\boldsymbol{D}_j + \boldsymbol{A}_i^{-1}\Big)\Big(\boldsymbol{z}_i - \boldsymbol{v}_{ij}\Big)\Big\}d\boldsymbol{z}_i,$$

which leads us to conclude that the solution to the integration is the inverse of the normalizing constant from a normal distribution with mean $\boldsymbol{v}_{ij}$ and covariance $(2\boldsymbol{D}_j + \boldsymbol{A}_i^{-1})^{-1}$:

$$\int \exp\Big\{-\frac{1}{2}\Big(\boldsymbol{z}_i - \boldsymbol{v}_{ij}\Big)^\top\Big(2\boldsymbol{D}_j + \boldsymbol{A}_i^{-1}\Big)\Big(\boldsymbol{z}_i - \boldsymbol{v}_{ij}\Big)\Big\}d\boldsymbol{z}_i = (2\pi)^{\frac{n}{2}}\det\Big(2\boldsymbol{D}_j + \boldsymbol{A}_i^{-1}\Big)^{-\frac{1}{2}} \propto \det\Big(2\boldsymbol{D}_j + \boldsymbol{A}_i^{-1}\Big)^{-\frac{1}{2}}.$$

Since $2\pi^{\frac{n}{2}}$ cancels with the same term from the distribution of the latent variables, this results in the final solution for the expectation:

$$\mathrm{E}\Big\{\exp\Big(\eta_{ij}\Big)\Big\} = \exp\Big[C_{ij} + \frac{1}{2}\Big\{\boldsymbol{v}_{ij}^\top\Big(2\boldsymbol{D}_j + \boldsymbol{A}_i^{-1}\Big)\boldsymbol{v}_{ij} - \boldsymbol{a}_i^\top \boldsymbol{A}_i^{-1}\boldsymbol{a}_i\Big\}\Big]\det\Big(2\boldsymbol{D}_j + \boldsymbol{A}_i^{-1}\Big)^{-\frac{1}{2}}\det\Big(\boldsymbol{A}_i\Big)^{-\frac{1}{2}}.$$

Nothing that this solution can be further simplified by working out the term inside the exponent, provides the final solution for the Poisson VA log-likelihood with the quadratic response model:

$$\begin{aligned}\log \mathcal{L}_{VA}(\boldsymbol{\Theta}, \boldsymbol{\xi}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\Big[y_{ij}\Big\{C_{ij} + \boldsymbol{a}_i^\top\boldsymbol{\gamma}_j - \boldsymbol{a}_i^\top\boldsymbol{D}_j\boldsymbol{a}_i - \mathrm{tr}\Big(\boldsymbol{D}_j\boldsymbol{A}_i\Big)\Big\} \\ - \exp\Big\{C_{ij} + \frac{1}{2}\Big((\boldsymbol{\gamma}_j + \boldsymbol{A}_i^{-1}\boldsymbol{a}_i)^\top(2\boldsymbol{D}_j + \boldsymbol{A}_i^{-1})^{-1}(\boldsymbol{\gamma}_j + \boldsymbol{A}_i^{-1}\boldsymbol{a}_i) \\ -\boldsymbol{a}_i^\top\boldsymbol{A}_i^{-1}\boldsymbol{a}_i\Big\}\det\Big\{2\boldsymbol{D}_j + \boldsymbol{A}_i^{-1}\Big\}^{-\frac{1}{2}}\det\Big\{\boldsymbol{A}_i\Big\}^{-\frac{1}{2}}\Big] + \frac{1}{2}\sum_{i=1}^{n}\Big\{\log\det\Big(\boldsymbol{A}_i\Big) - \mathrm{tr}\Big(\boldsymbol{A}_i\Big) - \boldsymbol{a}_i^\top\boldsymbol{a}_i\Big\}.\end{aligned} \tag{20}$$

Negative-Binomial: overdispersed counted responses

We model overdispersed counts using a Poisson-Gamma mixture distribution with a log-link function, where the Poisson rate parameter follows a Gamma distribution, resulting in a negative-binomial distribution, as

in Hui et al. (2017), where $y_{ij} \sim \mathcal{P}ois(\nu_{ij})$, with $\nu_{ij} \sim \Gamma(\phi_j, \phi_j/\mu_{ij})$, $\mathrm{E}(y_{ij}) = \mu_{ij}$ and $\mathrm{var}(y_{ij}) = \mu_{ij} + \mu_{ij}^2\phi_j$. This choice facilitates a closed form solution to the required integration, unlike other parametrizations of the negative-binomial distribution. A GLLVM with quadratic response model should only require a negative-binomial distribution for extremely overdispersed count data, i.e. when there is overdispersion for species that also have a narrow niche.

This choice requires finding an expression for $\mathrm{E}\{\exp(-\eta)\}$, which has a similar solution to $\mathrm{E}\{\exp(\eta)\}$, provided in the case of Poisson responses above. However, the calculation here includes the terms $(\boldsymbol{A}_i - 2\boldsymbol{D}_j)^{-1}$, which is required to be positive semi-definite or negative semi-definite, to determine a closed form solution to the integration. This assumption only fails when the resulting matrix is singular, which in practice happens when the off-diagonals of $\boldsymbol{A}_i$ are zero, and the diagonals of the matrices match, which should rarely be the case in practice.

The log-likelihood of negative-binomial distributed responses is:

$$\mathcal{L}(\boldsymbol{\Theta}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\left\{\left(y_{ij}+\phi_j-1\right)\log\left(\nu_{ij}\right)-\left(1+\frac{\phi_j}{\mu_{ij}}\right)\nu_{ij}+\phi(j)\left(\frac{\phi_j}{\mu_{ij}}\right)-\log\Gamma\left(\phi_j\right)-\log\left(y_{ij}!\right)\right\}-\frac{1}{2}\sum_{i=1}^{n}\boldsymbol{z}_i^\top\boldsymbol{z}_i, \tag{21}$$

where terms constant with respect to the parameters have been omitted. The optimal variational distribution for the latent variable ν_{ij} is:

$$\begin{aligned}
\log\{\mathrm{q}(\nu_{ij})\} &\propto \mathrm{E}\{\mathcal{L}(\boldsymbol{\Theta})\} \\
&\propto (y_{ij}+\phi_j-1)\log(\nu_{ij}) - [1+\phi_j\mathrm{E}\{\exp(-\eta_{ij})\}]\nu_{ij} \\
&\propto \left(y_{ij}+\phi_j-1\right)\log\left(\nu_{ij}\right) - \left[1+\phi_j\exp\left\{-C_{ij}\right.\right. \\
&\quad \left.\left.+\frac{1}{2}\left((-\boldsymbol{\gamma}_j+\boldsymbol{A}_i^{-1}\boldsymbol{a}_i)^\top(\boldsymbol{A}_i^{-1}-2\boldsymbol{D}_j)^{-1}(-\boldsymbol{\gamma}_j+\boldsymbol{A}_i^{-1}\boldsymbol{a}_i)-\boldsymbol{a}_i^\top\boldsymbol{A}_i^{-1}\boldsymbol{a}_i\right)\right\}\det\left(\boldsymbol{A}_i^{-1}-2\boldsymbol{D}_j\right)^{-\frac{1}{2}}\det\left(\boldsymbol{A}_i\right)^{-\frac{1}{2}}\right]\nu_{ij},
\end{aligned}$$

where terms constant with respect to the parameters have been omitted. Note, that the term $\det(\boldsymbol{A}_i^{-1} - 2\boldsymbol{D}_j)^{-\frac{1}{2}}$ needs to be calculated as $\boldsymbol{B}_{ij} = \boldsymbol{A}_i^{-1} - 2\boldsymbol{D}_j$ with $\boldsymbol{L}_{ij} = \mathrm{chol}(\boldsymbol{B}_{ij})$, so that $\det(\boldsymbol{A}_i^{-1} - 2\boldsymbol{D}_j)^{-\frac{1}{2}} = \det(\boldsymbol{L}_{ij})^{-1}$, where $\mathrm{chol}(\cdot)$ represents the cholesky factorization, to prevent having to potentially calculate the (undefined) square root of a negative determinant. From this we conclude that $\mathrm{q}(v_{ij}) \sim \Gamma(y_{ij}+\phi_j, \zeta_{ij})$, with:

$$\zeta_{ij} = 1+\phi_j\exp\left[-C_{ij}+\frac{1}{2}\left\{\left(-\boldsymbol{\gamma}_j+\boldsymbol{A}_i^{-1}\boldsymbol{a}_i\right)^\top\boldsymbol{B}^{-1}\left(-\boldsymbol{\gamma}_j+\boldsymbol{A}_i^{-1}\boldsymbol{a}_i\right)-\boldsymbol{a}_i^\top\boldsymbol{A}_i^{-1}\boldsymbol{a}_i\right\}\right]\det\left(\boldsymbol{L}_{ij}\right)^{-1}\det\left(\boldsymbol{A}_i\right)^{-\frac{1}{2}}.$$

The VA log-likelihood is derived by working out explicit expressions for the expectations in equation (13):

$$\begin{aligned}\mathrm{E}[\mathcal{L}\{\boldsymbol{\Theta}\}] = \sum_{i=1}^{n}\sum_{j=1}^{p}\Big[\Big\{y_{ij}+\phi_j-1\Big\}\Big\{\psi\Big(y_{ij}+\phi_j\Big)-\log\Big(\zeta_{ij}+\phi_j\Big)\Big\}-\Big\{y_{ij}+\phi_j\Big\}-\phi_j\tilde{\eta}_{ij}+\phi_j\log\Big\{\phi_j\Big\}-\log\Gamma\Big\{\phi_j\Big\}\Big] \\ -\frac{1}{2}\sum_{i=1}^{n}\Big\{\mathrm{tr}\Big(\boldsymbol{A}_i\Big)+\boldsymbol{a}_i^\top\boldsymbol{a}_i\Big\} \\ \mathrm{E}[\log\{\mathrm{q}(\nu_{ij})\}] = \sum_{i=1}^{n}\sum_{j=1}^{p}\Big[\Big\{y_{ij}+\phi_j-1\Big\}\Big\{\psi\Big(y_{ij}+\phi_j\Big)-\log\Big(\zeta_{ij}+\phi_j\Big)\Big\}-\Big(y_{ij}+\phi_j\Big)+\Big(y_{ij}+\phi_j\Big)\log\Big(\zeta_{ij}+\phi_j\Big) \\ -\log\Gamma\Big(y_{ij}+\phi_j\Big)\Big] \\ \mathrm{E}[\log\{\mathrm{q}(\boldsymbol{z}_i|\boldsymbol{a}_i,\boldsymbol{A}_i)\}] \propto -\frac{1}{2}\log\det(\boldsymbol{A}_i).\end{aligned}$$

Thus, VA log-likelihood for overdispersed counts and the quadratic response model is:

$$\begin{aligned}\mathcal{L}_{VA}(\boldsymbol{\Theta},\boldsymbol{\xi}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\Big\{-\phi_j\tilde{\eta}_{ij}-\Big(y_{ij}+\phi_j\Big)\log\Big(\zeta_{ij}\Big)+\log\Gamma\Big(y_{ij}+\phi_j\Big)+\phi_j\log\Big(\phi_j\Big)-\log\Gamma\Big(\phi_j\Big)\Big\} \\ +\frac{1}{2}\sum_{i=1}^{n}\Big\{\log\det\Big(\boldsymbol{A}_i\Big)-\mathrm{tr}\Big(\boldsymbol{A}_i\Big)-\boldsymbol{a}_i^\top\boldsymbol{a}_i\Big\}.\end{aligned} \tag{22}$$

Gamma responses: positive continuous

The log-likelihood for gamma distributed responses with a log-link function, shape $\frac{1}{\phi_j}$ and scale $\mu_{ij}\phi_j$ is:

$$\mathcal{L}(\boldsymbol{\Theta}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\Big[-\frac{1}{\phi_j}\Big\{\eta_{ij}+\log\phi_j-\log\Big(y_{ij}\Big)\Big\}-y_{ij}\frac{1}{\phi_j}\exp\Big\{-\eta_{ij}\Big\}-\log\Big\{y_{ij}\Big\}-\log\Gamma\Big\{\frac{1}{\phi_j}\Big\}\Big]-\frac{1}{2}\sum_{i=1}^{n}\boldsymbol{z}_i^\top\boldsymbol{z}_i, \tag{23}$$

where terms constant with respect to the parameters have been omitted. The VA log-likelihood is derived by working out expressions for the expectations in equation (4) with respect to the variational distribution of the latent variables:

$$\begin{aligned}\mathrm{E}[\mathcal{L}\{\boldsymbol{\Theta}\}] = \sum_{i=1}^{n}\sum_{j=1}^{p}\Big[-\frac{1}{\phi_j}\Big\{\mathrm{E}\Big(\eta_{ij}\Big)+\log\phi_j-\log\Big(y_{ij}\Big)\Big\}-y_{ij}\frac{1}{\phi_j}\mathrm{E}\Big\{\exp\Big(-\eta_{ij}\Big)\Big\} \\ -\log\Big(y_{ij}\Big)-\log\Gamma\Big\{\frac{1}{\phi_j}\Big\}\Big]-\frac{1}{2}\sum_{i=1}^{n}\Big\{\mathrm{tr}\Big(\boldsymbol{A}_i\Big)+\boldsymbol{a}_i^\top\boldsymbol{a}_i\Big\} \\ \mathrm{E}[\log\{\mathrm{q}(\boldsymbol{z}_i|\boldsymbol{a}_i,\boldsymbol{A}_i)\}] \propto -\frac{1}{2}\log\det(\boldsymbol{A}_i).\end{aligned}$$

The result for $\mathrm{E}(\eta_{ij})$ has been given above for all distributions, and the result for $\mathrm{E}\{\exp(-\eta_{ij})\}$ in the derivation of the VA log-likelihood for the negative-binomial distribution. Thus, VA log-likelihood for gamma responses and the quadratic model is:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\Theta},\boldsymbol{\xi}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\Bigg[-\frac{1}{\phi_j}\left\{\eta_{ij}+\log\left(\phi_j\right)-\log\left(y_{ij}\right)\right\} - y_{ij}\frac{1}{\phi_j}\exp\Bigg\{-C_{ij} \\ +\frac{1}{2}\left((-\boldsymbol{\gamma}_j+\boldsymbol{A}_i^{-1}\boldsymbol{a}_i)^\top\boldsymbol{B}_{ij}^{-1}(-\boldsymbol{\gamma}_j+\boldsymbol{A}_i^{-1}\boldsymbol{a}_i)-\boldsymbol{a}_i^\top\boldsymbol{A}_i^{-1}\boldsymbol{a}_i\right)\Bigg\}\det\left\{\boldsymbol{L}_{ij}\right\}^{-1}\det\left\{\boldsymbol{A}_i\right\}^{-\frac{1}{2}} - \log\left\{y_{ij}\right\} \\ -\log\Gamma\left\{\frac{1}{\phi_j}\right\}\Bigg] + \frac{1}{2}\sum_{i=1}^{n}\log\det\left(\boldsymbol{A}_i\right) - \mathrm{tr}\left(\boldsymbol{A}_i\right) - \boldsymbol{a}_i^\top\boldsymbol{a}_i,\end{aligned} \tag{25}$$

with $\boldsymbol{B}_{ij} = \boldsymbol{A}_i^{-1} - 2\boldsymbol{D}_j$ and $\boldsymbol{L}_{ij} = \mathrm{chol}(\boldsymbol{B}_{ij})$.

Approximate confidence intervals

After fitting the GLLVM by VA, confidence intervals for the parameters can be calculated from the approximate standard errors from the Hessian matrix, as described by Hui et al. (2017). The confidence intervals of the quadratic coefficients can be used to determine if species exhibit quadratic response curves, but otherwise lack any intuitive ecological interpretation. Therefore, we propose applying the Delta method to calculate approximate confidence intervals for species optima $\boldsymbol{u}_j$, tolerances $\boldsymbol{t}_j$, maxima c_j, and gradient lengths $4G_{qq}^{\frac{1}{2}}$, i.e. the ecological quantities of interest. Calculation of these confidence intervals is straightforward, and provides the possibility to determine if species differ in their optima $\boldsymbol{u}_j$, maxima c_j, or tolerances $\boldsymbol{t}_j$.

As noted by Hui et al. (2017), approximate asymptotic standard errors for parameters can be retrieved from the Hessian matrix of the VA log-likelihood. Here, we provide a brief overview of the Delta method, that can be used for the calculation of approximate standard errors of derived parameters such as the species optima $\boldsymbol{u}_j$, tolerances $\boldsymbol{t}_j$, species maxima c_j, and gradient lengths. The multivariate Delta method states, that approximate standard errors can be calculated for a parameter θ that is a function $f(\cdot)$ of a set of other parameters, by the gradient of the function with respect to the original parameters $\nabla\mathrm{f}(\cdot)$, and the variance-covariance matrix $\boldsymbol{\Sigma}$ of the original parameters. For example, the variance of the species optima, is given by:

$$\mathrm{Var}\{\mathrm{f}(\theta)\} \approx \frac{1}{2}\nabla\mathrm{f}(\gamma_{jq}, D_{jqq})^\top\boldsymbol{\Sigma}\nabla\mathrm{f}(\gamma_{jq}, D_{jqq}),$$

where θ is the optimum of a species response curve for a latent variable, and $\nabla\mathrm{f}(\gamma_{jq}, D_{jqq})$ is the first derivative of θ with respect to the separate parameters γ_{jq} and D_{jqq}. Similarly, the approximate covariances can be calculated. Tolerances and gradient lengths are only a function of the quadratic coefficient(s), not of

the linear coefficients.

Appendix S3: Fitting process

In this appendix, details on stabilizing the fitting of a GLLVM with quadratic response model are included. We used Template Model Builder [TMB; Kristensen et al. (2016)] to retrieve analytical derivatives and more smoothly fit the models.

Fitting

Even with the use of VA, estimation of GLLVMs is particularly challenging due to multimodality of the log-likelihood function. This often results in numerical optimization techniques getting stuck in local minima. We largely adopt the approach developed by Niku et al. (2019), who found that fitting GLLVMs with diagonal VA covariance matrices before fitting the model with unstructured VA covariance matrices, tended to stabilize the estimation for a GLLVM with linear response model. So, to overcome the problem of multimodality, we recommend to: 1) fit a common tolerances model with diagonal VA covariance matrix, 2) fit a model with the same structure, but with unstructured VA covariance matrix, for which the final solution of 1) is used as initial values, after which 3) the estimated VA means are used as the initial values to fit the unequal tolerances model (for which the same procedure with diagonal VA covariance matrix is potentially repeated). In practice, we found this procedure to stabilize fitting of the proposed GLLVM (which is included as the default in the `gllvm` R-package. Alternatively, the GLLVM can be fitted with diagonal or unstructured VA covariance matrix immediately, thus providing different fitting algorithms. Initial values for the numerical optimization are generated following the same procedure implemented in Niku et al. (2019), with the most basic option to start the optimization with a combination of zeros for the coefficients and ones for the latent variables, or with randomly generated initial values, or with initial values generated by: 1) fitting a multivariate Generalized Linear Model, 2) retrieving the Dunn-Smyth residuals (Dunn and Smyth, 1996) of the model, and 3) performing factor analysis on those residuals to retrieve initial values for quantities related to the latent variables (i.e. VA means and coefficients for the linear term of the model). It is also possible to use the solution of a vanilla GLLVM as initial values, though this usually performed either similar or worse than the approach for initial values described by Niku et al. (2019). When using the solution of a GLLVM with linear response model as initial values, the options for that GLLVM can be tweaked independently to tweaking the algorithm for the GLLVM with quadratic response model (e.g. by running it multiple times with different sets of initial values), resulting in an even larger number of potential fitting algorithms. In general, initial values that are closer to the final solution serve to further reduce the possibility for the optimizer to get stuck in local minima. Fitting the model with different sets of initial values can help to assess if it is has properly converged.

Estimation of the quadratic coefficients can be started at a small constant, or on a solution that is maximized conditionally on the initial values. Maximizing conditionally on the initial values further reduces the possibility for the optimizer to get stuck in local minima, in scenarios where the initial values of other coefficients, and of the latent variables, are close to an optimal solution. This approach works especially well in combination with first fitting the species-common tolerances model. It is possible to assess if a model converged to the solution that optimally maximizes the VA log-likelihood by examining the analytical gradient, in combination with the standard errors. The gradient should be close to zero if the model has converged. In our experience, the standard errors tend to be larger when a model has not (completely) converged.

If factor analysis cannot provide suitable initial values for the latent variables, Principal Component Analysis (PCA), Correspondence Analysis (CA), or Detrended Correspondence Analysis (DCA), can be considered instead. The software does accept externally generated initial values for the latent variables.

Regularisation

The quadratic model can be further extended with an additional species-common component, by changing the quadratic term to $\boldsymbol{z}_i^\top \left(\boldsymbol{D}_j + \boldsymbol{G}\right) \boldsymbol{z}_i$, where $\boldsymbol{G}$ is again a positive-definite diagonal matrix. Including a species-common component allows to make optimal use of the data, as parameters that are equal for all species can potentially be estimated with a higher degree of certainty than species-specific components $\boldsymbol{G}$. In this case, $\boldsymbol{D}_j$ accounts for species-specific deviation from the species-common components. The species-common components can be interpreted as an average measure of tolerance. This parametrization is unidentifiable, thus constraints for the species-specific component are required. For example, penalized likelihood methods could be used to enforce parameter identifiability (Tibshirani, 1996). This parametrization provides a hybrid form of the species-common and species-specific tolerances models, as with a large penalty it will fit the species-common tolerances model, and with a small or no penalty the species-specific tolerances model. This can additionally serve to further stabilize the fitting process (as discussed above and in Yee, 2004). Alternatively, the elements of $\boldsymbol{D}_j$ can be treated as a random effect, independent for species p and latent variables d, following a half-normal distribution, although we leave this as an avenue of future research.

Appendix S3: Code for simulations

Packages required for simulations

Load required libraries

- gllvm to run the model (at moment of writing this needs to be retrieved from github)
- vegan for procrustes
- parallel and foreach for parallel processing

```
library(gllvm)
library(vegan)
library(foreach)
library(parallel)
library(doSNOW)
```

Simulate data

This chunk is the same for all distributions.

```
# procrustes function that excludes values larger than 10, for optima.
# Edited from the vegan package.
procrustes.gllvm <- function(X, Y, symmetric = TRUE, scale = TRUE, threshold = 10,
    ...) {
    if (nrow(X) != nrow(Y))
        stop(gettextf("matrices have different number of rows: %d and %d",
            nrow(X), nrow(Y)))
    if (ncol(X) != ncol(Y))
        stop(gettextf("matrices have different number of columns: %d and %d",
            ncol(X), ncol(Y)))

    idx <- apply(X, 2, function(i) (!(i > threshold | i < (-threshold))))
    idx[upper.tri(idx)] <- F  #for identifiability constraint
    ctrace <- function(MAT, idx) {
```

```
    sum(MAT[idx]^2)
}
c <- 1
if (symmetric) {
    X <- t(t(X) - sapply(1:ncol(X), function(i) mean(X[idx[, i], i])))
    Y <- t(t(Y) - sapply(1:ncol(X), function(i) mean(Y[idx[, i], i])))
    X <- X/sqrt(ctrace(X, idx))
    Y <- Y/sqrt(ctrace(Y, idx))
}
xmean <- sapply(1:ncol(X), function(i) mean(X[idx[, i], i]))
ymean <- sapply(1:ncol(X), function(i) mean(Y[idx[, i], i]))
if (!symmetric) {
    X <- t(t(X) - sapply(1:ncol(X), function(i) mean(X[idx[, i], i])))
    Y <- t(t(Y) - sapply(1:ncol(X), function(i) mean(Y[idx[, i], i])))
}
XY <- matrix(0, ncol = ncol(X), nrow = ncol(X))
for (i in 1:ncol(X)) {
    for (j in 1:ncol(X)) {
        XY[i, j] <- sum(X[idx[, i], i] * Y[idx[, i], j])
    }
}

sol <- svd(XY)
A <- sol$v %*% t(sol$u)
if (scale) {
    c <- sum(sol$d)/ctrace(Y)
}
Yrot <- c * Y %*% A

b <- xmean - c * ymean %*% A
R2 <- ctrace(X, idx) + c * c * ctrace(Y, idx) - 2 * c * sum(sol$d)
reslt <- list(Yrot = Yrot, X = X, ss = R2, rotation = A, translation = b,
```

```
        scale = c, xmean = xmean, symmetric = symmetric, call = match.call())
    reslt$svd <- sol
    class(reslt) <- "procrustes"
    reslt
}



get_eta <- function(n, p) {
    x <- mvtnorm::rmvnorm(n, rep(0, 2), diag(2))
    opt1 <- runif(p, range(x[, 1])[1], range(x[, 1])[2])
    opt2 <- runif(p, range(x[, 2])[1], range(x[, 2])[2])
    opt <- cbind(opt1, opt2)

    opt[upper.tri(opt, diag = F)] <- 0
    tol <- matrix(runif(p * 2, 0.2, 1), ncol = 2)
    c <- runif(p, 2, 6)
    linpred <- matrix(c, ncol = p, nrow = n, byrow = T) - matrix(rowSums(opt^2/(2 *
        tol^2)), ncol = p, nrow = n, byrow = T) - x^2 %*% t(1/((2 * tol^2))) +
        2 * x %*% t(opt/(2 * tol^2))

    return(list(eta = linpred, opt = opt, tol = tol, lv = x, max = c))
}



# function to simulate data
sim_dat <- function(eta, family, seed, phi = NULL, cutoffs = NULL) {
    n <- dim(eta)[1]
    p <- dim(eta)[2]
    if (is.null(phi))
        phi <- rep(1, p)
    if (is.null(cutoffs))
        cutoffs <- 1:5
```

```r
if (!(family %in% c("gaussian", "poisson", "negative.binomial", "binomial",
    "ordinal", "gamma"))) {
    stop("Wrong family.")
}
if (family == "gaussian") {
    if (length(phi) != p)
        stop("Wrong length phi supplied.")
    set.seed(seed)
    mu <- eta
    y <- matrix(NA, n, p)

    for (j in 1:p) {
        y[, j] <- rnorm(n, mean = mu[, j], sd = phi[j])
    }

} else if (family == "poisson") {
    set.seed(seed)
    mu <- exp(eta)
    y <- matrix(NA, n, p)

    for (j in 1:p) {
        y[, j] <- rpois(n, mu[, j])
    }

}
if (family == "negative.binomial") {
    if (length(phi) != p)
        stop("Wrong length phi supplied.")
    set.seed(seed)
    mu <- exp(eta)
    y <- matrix(NA, n, p)
```

```
    for (j in 1:p) {
        y[, j] <- rnbinom(n, mu = mu[, j], size = mu[, j] * phi[j])
    }
}
if (family == "binomial") {
    set.seed(seed)
    mu <- pnorm(eta)
    y <- matrix(NA, n, p)

    for (j in 1:p) {
        y[, j] <- rbinom(n = n, size = 1, prob = mu[, j])
    }
}
if (family == "ordinal") {
    set.seed(seed)

    k.max <- length(cutoffs) + 1
    preds <- array(NA, dim = c(k.max, n, p), dimnames = list(paste("level",
        1:max(k.max), sep = ""), NULL, NULL))

    for (i in 1:n) {
        for (j in 1:p) {
            probK <- NULL
            probK[1] <- pnorm(cutoffs[1] - eta[i, j], log.p = FALSE)
            probK[k.max] <- 1 - pnorm(cutoffs[k.max - 1] - eta[i, j])
            levels <- 2:(k.max - 1)
            for (k in levels) {
              probK[k] <- pnorm(cutoffs[k] - eta[i, j]) - pnorm(cutoffs[k -
                1] - eta[i, j])
            }
            preds[, i, j] <- c(probK)
        }
```

```
        }
        y = matrix(NA, nrow = n, ncol = p)
        k <- length(cutoffs) + 1
        for (j in 1:p) {
            for (i in 1:n) {
                y[i, j] <- sample(k, 1, prob = preds[, i, j][!is.na(preds[,
                  i, j])])
            }
        }
    } else if (family == "gamma") {
        if (length(phi) != p)
            stop("Wrong length phi supplied.")
        set.seed(seed)
        mu <- exp(eta)
        y <- matrix(NA, n, p)

        for (j in 1:p) {
            y[, j] <- rgamma(n, shape = 1/phi[j], scale = phi[j] * mu[,
                j])
        }
    }
    return(y)

}

# function to fit models and run simulation
sim_gllvm <- function(n, p, r = 1000, family) {

    progress_n <- function(i) cat(sprintf(paste(paste("task %d of", r),
        "is now complete\n"), i))
    opts_n <- list(progress = progress_n)
```

```r
set.seed(1)
sim <- get_eta(n, p)
eta <- sim$eta
true_opt <- sim$opt
true_lv <- sim$lv

result <- foreach(i = 1:r, .inorder = F, .packages = c("gllvm", "vegan"),
    .export = c("sim_dat", "procrustes.gllvm"), .options.snow = opts_n) %dopar%
    {

        dat <- sim_dat(eta, family, seed = i, nsim = r)
        dat2 <- dat

        # Make sure that classes are sequential
        if (family == "ordinal" & any(diff(sort(unique(c(dat)))) !=
            1)) {
            while (any(diff(sort(unique(c(dat)))) != 1)) {
              maxK <- max(dat)
              dat[dat == maxK] <- length(unique(c(dat)))  #This might be wrong, double check
            }
        }
        idx <- colSums(dat) != 0
        idx2 <- rowSums(dat) != 0
        dat <- dat[idx2, idx]

        mod <- list()
        mod <- try(gllvm(dat, num.lv = 2, family = family, quadratic = TRUE,
            control = list(maxit = 1e+06), control.start = list(starting.val = "res",
              start.struc = "LV", n.init = 1), control.va = list(diag.iter = 1),
            sd.errors = F, zeta.struc = "common"), silent = T)
        if (!inherits(mod, "try-error")) {
            mod$start <- "res_LV_diagiter"
```

```
}
if (inherits(mod, "try-error")) {
    mod <- try(gllvm(dat, num.lv = 2, family = family, quadratic = TRUE,
      control = list(maxit = 1e+06), control.start = list(starting.val = "res",
        start.struc = "LV", n.init = 1), control.va = list(diag.iter = 0),
      sd.errors = F, zeta.struc = "common"), silent = T)
    if (!inherits(mod, "try-error")) {
      mod$start <- "res_LV"
    }
} else if (is.infinite(logLik(mod))) {
    mod <- try(gllvm(dat, num.lv = 2, family = family, quadratic = TRUE,
      control = list(maxit = 1e+06), control.start = list(starting.val = "res",
        start.struc = "LV", n.init = 1), control.va = list(diag.iter = 0),
      sd.errors = F, zeta.struc = "common"), silent = T)
    if (!inherits(mod, "try-error")) {
      mod$start <- "res_LV"
    }
}
if (inherits(mod, "try-error")) {
    mod <- try(gllvm(dat, num.lv = 2, family = family, quadratic = TRUE,
      control = list(maxit = 1e+06), control.start = list(starting.val = "res",
        start.struc = "all", n.init = 1), control.va = list(diag.iter = 1),
      sd.errors = F, zeta.struc = "common"), silent = T)
    if (!inherits(mod, "try-error")) {
      mod$start <- "res_all_diagiter"
    }
} else if (is.infinite(logLik(mod))) {
    mod <- try(gllvm(dat, num.lv = 2, family = family, quadratic = TRUE,
      control = list(maxit = 1e+06), control.start = list(starting.val = "res",
        start.struc = "all", n.init = 1), control.va = list(diag.iter = 1),
      sd.errors = F, zeta.struc = "common"), silent = T)
    if (!inherits(mod, "try-error")) {
```

```
      mod$start <- "res_all_diagiter"
    }
  }
  if (inherits(mod, "try-error")) {
    mod <- try(gllvm(dat, num.lv = 2, family = family, quadratic = TRUE,
      control = list(maxit = 1e+06), control.start = list(starting.val = "res",
        start.struc = "all", n.init = 1), control.va = list(diag.iter = 0),
      sd.errors = F, zeta.struc = "common"), silent = T)
    if (!inherits(mod, "try-error")) {
      mod$start <- "res_all"
    }
  } else if (is.infinite(logLik(mod))) {
    mod <- try(gllvm(dat, num.lv = 2, family = family, quadratic = TRUE,
      control = list(maxit = 1e+06), control.start = list(starting.val = "res",
        start.struc = "all", n.init = 1), control.va = list(diag.iter = 0),
      sd.errors = F, zeta.struc = "common"), silent = T)
    if (!inherits(mod, "try-error")) {
      mod$start <- "res_all"
    }
  }

  if (inherits(mod, "try-error")) {
    mod <- try(gllvm(dat, num.lv = 2, family = family, quadratic = TRUE,
      control = list(maxit = 1e+06), control.start = list(starting.val = "zero",
        start.struc = "LV", n.init = 1), control.va = list(diag.iter = 1),
      sd.errors = F, zeta.struc = "common"), silent = T)
    if (!inherits(mod, "try-error")) {
      mod$start <- "zero_LV_diagiter"
    }
  } else if (is.infinite(logLik(mod))) {
    mod <- try(gllvm(dat, num.lv = 2, family = family, quadratic = TRUE,
      control = list(maxit = 1e+06), control.start = list(starting.val = "zero",
```

```
      start.struc = "LV", n.init = 1), control.va = list(diag.iter = 1),
    sd.errors = F, zeta.struc = "common"), silent = T)
  if (!inherits(mod, "try-error")) {
    mod$start <- "zero_LV_diagiter"
  }
}

if (inherits(mod, "try-error")) {
  mod <- try(gllvm(dat, num.lv = 2, family = family, quadratic = TRUE,
    control = list(maxit = 1e+06), control.start = list(starting.val = "zero",
      start.struc = "LV", n.init = 1), control.va = list(diag.iter = 0),
    sd.errors = F, zeta.struc = "common"), silent = T)
  if (!inherits(mod, "try-error")) {
    mod$start <- "zero_LV"
  }
} else if (is.infinite(logLik(mod))) {
  mod <- try(gllvm(dat, num.lv = 2, family = family, quadratic = TRUE,
    control = list(maxit = 1e+06), control.start = list(starting.val = "zero",
      start.struc = "LV", n.init = 1), control.va = list(diag.iter = 0),
    sd.errors = F, zeta.struc = "common"), silent = T)
  if (!inherits(mod, "try-error")) {
    mod$start <- "zero_LV"
  }
}

if (inherits(mod, "try-error")) {
  mod <- try(gllvm(dat, num.lv = 2, family = family, quadratic = TRUE,
    control = list(maxit = 1e+06), control.start = list(starting.val = "zero",
      start.struc = "all", n.init = 1), control.va = list(diag.iter = 1),
    sd.errors = F, zeta.struc = "common"), silent = T)
  if (!inherits(mod, "try-error")) {
    mod$start <- "zero_all_diagiter"
```

```
    }
} else if (is.infinite(logLik(mod))) {
    mod <- try(gllvm(dat, num.lv = 2, family = family, quadratic = TRUE,
      control = list(maxit = 1e+06), control.start = list(starting.val = "zero",
        start.struc = "all", n.init = 1), control.va = list(diag.iter = 1),
      sd.errors = F, zeta.struc = "common"), silent = T)
    if (!inherits(mod, "try-error")) {
      mod$start <- "zero_all_diagiter"
    }
}

if (inherits(mod, "try-error")) {
    mod <- try(gllvm(dat, num.lv = 2, family = family, quadratic = TRUE,
      control = list(maxit = 1e+06), control.start = list(starting.val = "zero",
        start.struc = "all", n.init = 1), control.va = list(diag.iter = 0),
      sd.errors = F, zeta.struc = "common"), silent = T)
    if (!inherits(mod, "try-error")) {
      mod$start <- "zero_LV"
    }
} else if (is.infinite(logLik(mod))) {
    mod <- try(gllvm(dat, num.lv = 2, family = family, quadratic = TRUE,
      control = list(maxit = 1e+06), control.start = list(starting.val = "zero",
        start.struc = "all", n.init = 1), control.va = list(diag.iter = 0),
      sd.errors = F, zeta.struc = "common"), silent = T)
    if (!inherits(mod, "try-error")) {
      mod$start <- "zero_LV"
    }
}

# extra insurance to make sure we don't end up with a common tolerances
# model
if (!inherits(mod, "try-error")) {
```

```
if (length(mod$params$theta[, 3:4][!duplicated(round(mod$params$theta[,
  3:4], 2))]) == 2) {
  mod <- try(gllvm(dat, num.lv = 2, family = family, quadratic = TRUE,
    control = list(maxit = 1e+06), control.start = list(starting.val = "res",
      start.struc = "all", n.init = 1), control.va = list(diag.iter = 1),
    sd.errors = F, zeta.struc = "common"), silent = T)
  if (!inherits(mod, "try-error")) {
    mod$start <- "res_all_diagiter"
  }
  if (inherits(mod, "try-error")) {
    mod <- try(gllvm(dat, num.lv = 2, family = family,
      quadratic = TRUE, control = list(maxit = 1e+06),
      control.start = list(starting.val = "res", start.struc = "all",
        n.init = 1), control.va = list(diag.iter = 0),
      sd.errors = F, zeta.struc = "common"), silent = T)
    if (!inherits(mod, "try-error")) {
      mod$start <- "res_all"
    }
  } else if (is.infinite(logLik(mod))) {
    mod <- try(gllvm(dat, num.lv = 2, family = family,
      quadratic = TRUE, control = list(maxit = 1e+06),
      control.start = list(starting.val = "res", start.struc = "all",
        n.init = 1), control.va = list(diag.iter = 0),
      sd.errors = F, zeta.struc = "common"), silent = T)
    if (!inherits(mod, "try-error")) {
      mod$start <- "res_all"
    }
  }

  if (inherits(mod, "try-error")) {
    mod <- try(gllvm(dat, num.lv = 2, family = family,
      quadratic = TRUE, control = list(maxit = 1e+06),
```

```
    control.start = list(starting.val = "zero", start.struc = "all",
      n.init = 1), control.va = list(diag.iter = 1),
    sd.errors = F, zeta.struc = "common"), silent = T)
  if (!inherits(mod, "try-error")) {
    mod$start <- "zero_all_diagiter"
  }
} else if (is.infinite(logLik(mod))) {
  mod <- try(gllvm(dat, num.lv = 2, family = family,
    quadratic = TRUE, control = list(maxit = 1e+06),
    control.start = list(starting.val = "zero", start.struc = "all",
      n.init = 1), control.va = list(diag.iter = 1),
    sd.errors = F, zeta.struc = "common"), silent = T)
  if (!inherits(mod, "try-error")) {
    mod$start <- "zero_all_diagiter"
  }
}

if (inherits(mod, "try-error")) {
  mod <- try(gllvm(dat, num.lv = 2, family = family,
    quadratic = TRUE, control = list(maxit = 1e+06),
    control.start = list(starting.val = "zero", start.struc = "all",
      n.init = 1), control.va = list(diag.iter = 0),
    sd.errors = F, zeta.struc = "common"), silent = T)
  if (!inherits(mod, "try-error")) {
    mod$start <- "zero_LV"
  }
} else if (is.infinite(logLik(mod))) {
  mod <- try(gllvm(dat, num.lv = 2, family = family,
    quadratic = TRUE, control = list(maxit = 1e+06),
    control.start = list(starting.val = "zero", start.struc = "all",
      n.init = 1), control.va = list(diag.iter = 0),
    sd.errors = F, zeta.struc = "common"), silent = T)
```

```
            if (!inherits(mod, "try-error")) {
              mod$start <- "zero_LV"
            }
          }

      }
  }
  result <- list()
  if (!inherits(mod, "try-error")) {
      # exclude species without observations if any
      theta <- mod$params$theta
      opt <- theta[, 1:2]/(2 * abs(theta[, 3:4]))
      true_opt2 <- true_opt[idx, ]

      X <- opt
      Y <- true_opt2
      idx3 <- (X > 10 | X < (-10))
      idx3[upper.tri(idx3)] <- T
      X <- mod$lvs
      Y <- true_lv[idx2, ]

      result[[1]] <- mod$params$theta
      result[[2]] <- mod$lvs
      result[[3]] <- dat2
      result[[4]] <- mod$call
      result[[5]] <- mod$start
      result[[6]] <- try(procrustes.gllvm(mod$lvs, true_lv[idx2,
        ], symmetric = T)$ss)
      result[[7]] <- try(procrustes.gllvm(opt, true_opt2, symmetric = T)$ss)
      result[[8]] <- sum(idx3) - 1
      remove(mod)
  } else {
```

```
                result[[1]] <- result[[2]] <- result[[4]] <- result[[5]] <- result[[6]] <- result[[7]] <-
                result[[3]] <- dat2
            }

            return(result)

        }
    return(result)
}
```

Poisson

```
cores <- detectCores()
cl <- makeCluster(cores[1] - 1)
registerDoSNOW(cl)

# keeping n the same: high.
result1 <- sim_gllvm(100, 20, r = 1000, family = "poisson")
result2 <- sim_gllvm(100, 30, r = 1000, family = "poisson")
result3 <- sim_gllvm(100, 40, r = 1000, family = "poisson")
result4 <- sim_gllvm(100, 50, r = 1000, family = "poisson")
result5 <- sim_gllvm(100, 60, r = 1000, family = "poisson")
result6 <- sim_gllvm(100, 70, r = 1000, family = "poisson")
result7 <- sim_gllvm(100, 80, r = 1000, family = "poisson")
result8 <- sim_gllvm(100, 90, r = 1000, family = "poisson")
result9 <- sim_gllvm(100, 100, r = 1000, family = "poisson")
result <- list(result1, result2, result3, result4, result5, result6, result7,
    result8, result9)

# keeping p the same: high.
result1 <- sim_gllvm(20, 100, r = 1000, family = "poisson")
result2 <- sim_gllvm(30, 100, r = 1000, family = "poisson")
```

```
result3 <- sim_gllvm(40, 100, r = 1000, family = "poisson")
result4 <- sim_gllvm(50, 100, r = 1000, family = "poisson")
result5 <- sim_gllvm(60, 100, r = 1000, family = "poisson")
result6 <- sim_gllvm(70, 100, r = 1000, family = "poisson")
result7 <- sim_gllvm(80, 100, r = 1000, family = "poisson")
result8 <- sim_gllvm(90, 100, r = 1000, family = "poisson")
result2 <- list(result1, result2, result3, result4, result5, result6, result7,
    result8, result9)
```

Binomial

```
cores <- detectCores()
cl <- makeCluster(cores[1] - 1)
registerDoSNOW(cl)

# keeping n the same: high.
result1 <- sim_gllvm(100, 20, r = 1000, family = "binomial")
result2 <- sim_gllvm(100, 30, r = 1000, family = "binomial")
result3 <- sim_gllvm(100, 40, r = 1000, family = "binomial")
result4 <- sim_gllvm(100, 50, r = 1000, family = "binomial")
result5 <- sim_gllvm(100, 60, r = 1000, family = "binomial")
result6 <- sim_gllvm(100, 70, r = 1000, family = "binomial")
result7 <- sim_gllvm(100, 80, r = 1000, family = "binomial")
result8 <- sim_gllvm(100, 90, r = 1000, family = "binomial")
result9 <- sim_gllvm(100, 100, r = 1000, family = "binomial")
result <- list(result1, result2, result3, result4, result5, result6, result7,
    result8, result9)

# keeping p the same: high.
result1 <- sim_gllvm(20, 100, r = 1000, family = "binomial")
result2 <- sim_gllvm(30, 100, r = 1000, family = "binomial")
result3 <- sim_gllvm(40, 100, r = 1000, family = "binomial")
```

```
result4 <- sim_gllvm(50, 100, r = 1000, family = "binomial")
result5 <- sim_gllvm(60, 100, r = 1000, family = "binomial")
result6 <- sim_gllvm(70, 100, r = 1000, family = "binomial")
result7 <- sim_gllvm(80, 100, r = 1000, family = "binomial")
result8 <- sim_gllvm(90, 100, r = 1000, family = "binomial")
result2 <- list(result1, result2, result3, result4, result5, result6, result7,
    result8, result9)
```

Negative-Binomial

```
cores <- detectCores()
cl <- makeCluster(cores[1] - 1)
registerDoSNOW(cl)

# keeping n the same: high.
result1 <- sim_gllvm(100, 20, r = 1000, family = "negative.binomial")
result2 <- sim_gllvm(100, 30, r = 1000, family = "negative.binomial")
result3 <- sim_gllvm(100, 40, r = 1000, family = "negative.binomial")
result4 <- sim_gllvm(100, 50, r = 1000, family = "negative.binomial")
result5 <- sim_gllvm(100, 60, r = 1000, family = "negative.binomial")
result6 <- sim_gllvm(100, 70, r = 1000, family = "negative.binomial")
result7 <- sim_gllvm(100, 80, r = 1000, family = "negative.binomial")
result8 <- sim_gllvm(100, 90, r = 1000, family = "negative.binomial")
result9 <- sim_gllvm(100, 100, r = 1000, family = "negative.binomial")
result <- list(result1, result2, result3, result4, result5, result6, result7,
    result8, result9)

# keeping p the same: high.
result1 <- sim_gllvm(20, 100, r = 1000, family = "negative.binomial")
result2 <- sim_gllvm(30, 100, r = 1000, family = "negative.binomial")
result3 <- sim_gllvm(40, 100, r = 1000, family = "negative.binomial")
result4 <- sim_gllvm(50, 100, r = 1000, family = "negative.binomial")
```

```
result5 <- sim_gllvm(60, 100, r = 1000, family = "negative.binomial")
result6 <- sim_gllvm(70, 100, r = 1000, family = "negative.binomial")
result7 <- sim_gllvm(80, 100, r = 1000, family = "negative.binomial")
result8 <- sim_gllvm(90, 100, r = 1000, family = "negative.binomial")
result2 <- list(result1, result2, result3, result4, result5, result6, result7,
    result8, result9)
```

Ordinal

```
cores <- detectCores()
cl <- makeCluster(cores[1] - 1)
registerDoSNOW(cl)

# keeping n the same: high.
result1 <- sim_gllvm(100, 20, r = 1000, family = "ordinal")
result2 <- sim_gllvm(100, 30, r = 1000, family = "ordinal")
result3 <- sim_gllvm(100, 40, r = 1000, family = "ordinal")
result4 <- sim_gllvm(100, 50, r = 1000, family = "ordinal")
result5 <- sim_gllvm(100, 60, r = 1000, family = "ordinal")
result6 <- sim_gllvm(100, 70, r = 1000, family = "ordinal")
result7 <- sim_gllvm(100, 80, r = 1000, family = "ordinal")
result8 <- sim_gllvm(100, 90, r = 1000, family = "ordinal")
result9 <- sim_gllvm(100, 100, r = 1000, family = "ordinal")
result <- list(result1, result2, result3, result4, result5, result6, result7,
    result8, result9)

# keeping p the same: high.
result1 <- sim_gllvm(20, 100, r = 1000, family = "ordinal")
result2 <- sim_gllvm(30, 100, r = 1000, family = "ordinal")
result3 <- sim_gllvm(40, 100, r = 1000, family = "ordinal")
result4 <- sim_gllvm(50, 100, r = 1000, family = "ordinal")
result5 <- sim_gllvm(60, 100, r = 1000, family = "ordinal")
```

```
result6 <- sim_gllvm(70, 100, r = 1000, family = "ordinal")
result7 <- sim_gllvm(80, 100, r = 1000, family = "ordinal")
result8 <- sim_gllvm(90, 100, r = 1000, family = "ordinal")
result2 <- list(result1, result2, result3, result4, result5, result6, result7,
    result8, result9)
```

Gaussian

```
cores <- detectCores()
cl <- makeCluster(cores[1] - 1)
registerDoSNOW(cl)

# keeping n the same: high.
result1 <- sim_gllvm(100, 20, r = 1000, family = "gaussian")
result2 <- sim_gllvm(100, 30, r = 1000, family = "gaussian")
result3 <- sim_gllvm(100, 40, r = 1000, family = "gaussian")
result4 <- sim_gllvm(100, 50, r = 1000, family = "gaussian")
result5 <- sim_gllvm(100, 60, r = 1000, family = "gaussian")
result6 <- sim_gllvm(100, 70, r = 1000, family = "gaussian")
result7 <- sim_gllvm(100, 80, r = 1000, family = "gaussian")
result8 <- sim_gllvm(100, 90, r = 1000, family = "gaussian")
result9 <- sim_gllvm(100, 100, r = 1000, family = "gaussian")
result <- list(result1, result2, result3, result4, result5, result6, result7,
    result8, result9)

# keeping p the same: high.
result1 <- sim_gllvm(20, 100, r = 1000, family = "gaussian")
result2 <- sim_gllvm(30, 100, r = 1000, family = "gaussian")
result3 <- sim_gllvm(40, 100, r = 1000, family = "gaussian")
result4 <- sim_gllvm(50, 100, r = 1000, family = "gaussian")
result5 <- sim_gllvm(60, 100, r = 1000, family = "gaussian")
result6 <- sim_gllvm(70, 100, r = 1000, family = "gaussian")
```

```
result7 <- sim_gllvm(80, 100, r = 1000, family = "gaussian")
result8 <- sim_gllvm(90, 100, r = 1000, family = "gaussian")
result2 <- list(result1, result2, result3, result4, result5, result6, result7,
    result8, result9)
```

Gamma

```
cores <- detectCores()
cl <- makeCluster(cores[1] - 1)
registerDoSNOW(cl)

# keeping n the same: high.
result1 <- sim_gllvm(100, 20, r = 1000, family = "gamma")
result2 <- sim_gllvm(100, 30, r = 1000, family = "gamma")
result3 <- sim_gllvm(100, 40, r = 1000, family = "gamma")
result4 <- sim_gllvm(100, 50, r = 1000, family = "gamma")
result5 <- sim_gllvm(100, 60, r = 1000, family = "gamma")
result6 <- sim_gllvm(100, 70, r = 1000, family = "gamma")
result7 <- sim_gllvm(100, 80, r = 1000, family = "gamma")
result8 <- sim_gllvm(100, 90, r = 1000, family = "gamma")
result9 <- sim_gllvm(100, 100, r = 1000, family = "gamma")
result <- list(result1, result2, result3, result4, result5, result6, result7,
    result8, result9)

# keeping p the same: high.
result1 <- sim_gllvm(20, 100, r = 1000, family = "gamma")
result2 <- sim_gllvm(30, 100, r = 1000, family = "gamma")
result3 <- sim_gllvm(40, 100, r = 1000, family = "gamma")
result4 <- sim_gllvm(50, 100, r = 1000, family = "gamma")
result5 <- sim_gllvm(60, 100, r = 1000, family = "gamma")
result6 <- sim_gllvm(70, 100, r = 1000, family = "gamma")
result7 <- sim_gllvm(80, 100, r = 1000, family = "gamma")
```

```r
result8 <- sim_gllvm(90, 100, r = 1000, family = "gamma")

result2 <- list(result1, result2, result3, result4, result5, result6, result7,
    result8, result9)
```

Appendix S5: Extra results for real data examples

Distribution	num. LV	Quadratic?	Start.struc	Row.eff	df	logLik	AICc	delta AICc
Poisson	3	T			81.00	-626.05	1466.39	0.00
Poisson	2	T		fixed	86.00	-618.25	1468.59	2.20
Poisson	3	T	all		81.00	-638.27	1490.84	24.45
Poisson	3	T	all	random	82.00	-639.72	1497.25	30.86
Poisson	2	T			59.00	-679.40	1502.46	36.06
Poisson	2	T	all	random	60.00	-679.76	1506.14	39.74
Poisson	2	T	all		59.00	-699.85	1543.35	76.96
NB	1	T		fixed	75.00	-676.35	1546.54	80.15
Poisson	3	LV			48.00	-726.81	1566.00	99.61
Poisson	3	LV		random	49.00	-726.81	1568.75	102.35
NB	2	F			47.00	-733.68	1577.03	110.63
NB	2	F		random	48.00	-734.81	1582.01	115.62
NB	2	LV			49.00	-733.64	1582.41	116.01
NB	2	F		fixed	74.00	-696.05	1582.63	116.24
NB	1	F		fixed	63.00	-713.98	1583.60	117.21
NB	2	LV		random	50.00	-733.64	1585.17	118.77
Poisson	3	F		fixed	72.00	-701.20	1586.37	119.98
NB	1	LV		fixed	64.00	-713.98	1586.66	120.26
NB	2	LV		fixed	76.00	-695.54	1588.28	121.88
NB	1	LV			37.00	-757.81	1599.07	132.67
NB	1	F			36.00	-759.40	1599.71	133.32
NB	2	T			71.00	-709.88	1600.48	134.08
NB	1	LV		random	38.00	-757.81	1601.61	135.22
NB	1	F		random	37.00	-759.40	1602.24	135.85
NB	2	T		random	72.00	-709.88	1603.72	137.33
Poisson	3	F		random	46.00	-748.73	1604.42	138.03
NB	3	F			57.00	-733.68	1605.15	138.75
NB	3	F		random	58.00	-734.81	1610.33	143.93
NB	1	T			48.00	-749.21	1610.80	144.41
NB	1	T		random	49.00	-749.21	1613.55	147.15
NB	3	LV			60.00	-733.64	1613.89	147.50
NB	3	F		fixed	84.00	-696.05	1617.00	150.60
NB	3	LV		random	61.00	-733.85	1617.32	150.92
Poisson	3	F			45.00	-758.85	1621.97	155.58
NB	3	LV		fixed	87.00	-696.05	1627.85	161.45
Poisson	2	LV		fixed	64.00	-742.14	1642.98	176.59
NB	2	T		fixed	98.00	-684.74	1647.35	180.96
Poisson	2	F		fixed	62.00	-755.90	1664.42	198.02
NB	3	T			93.00	-709.88	1678.00	211.60
NB	3	T		random	94.00	-709.88	1681.86	215.46
Poisson	2	LV		random	38.00	-801.46	1688.89	222.50
Poisson	2	LV			37.00	-804.34	1692.12	225.72
Poisson	2	F		random	36.00	-809.97	1700.84	234.45
NB	3	T		fixed	120.00	-678.59	1732.25	265.86
Poisson	2	F			35.00	-845.83	1770.06	303.66
Poisson	1	T		fixed	63.00	-889.80	1935.25	468.86
Poisson	1	T	all	random	37.00	-1000.41	2084.26	617.87
Poisson	1	LV		fixed	52.00	-1107.98	2339.45	873.05
Poisson	1	F		fixed	51.00	-1118.82	2358.31	891.92
Poisson	1	LV		random	26.00	-1175.57	2407.68	941.28
Poisson	1	F		random	25.00	-1181.47	2417.13	950.74
Poisson	1	T	all		36.00	-1264.96	2610.83	1144.43
Poisson	1	T			36.00	-1265.04	2610.98	1144.59
Poisson	1	LV			25.00	-1402.81	2859.82	1393.42
Poisson	1	F			24.00	-1425.15	2902.16	1435.76
Poisson	3	LV		fixed	75.00	-1496.10	3186.05	1719.65
Poisson	3	T		fixed	108.00	-1496.10	3311.92	1845.52

Table S1: Results for GLLVMs with quadratic response model fitted to the hunting spider dataset. The column quadratic includes species-specific tolerances (TRUE), common-tolerances (LV), and no explicit quadratic response (FALSE). The column start.struc is applicable only to the species-specific tolerances model, and indicates if a common-tolerances model was not fitted first (all).

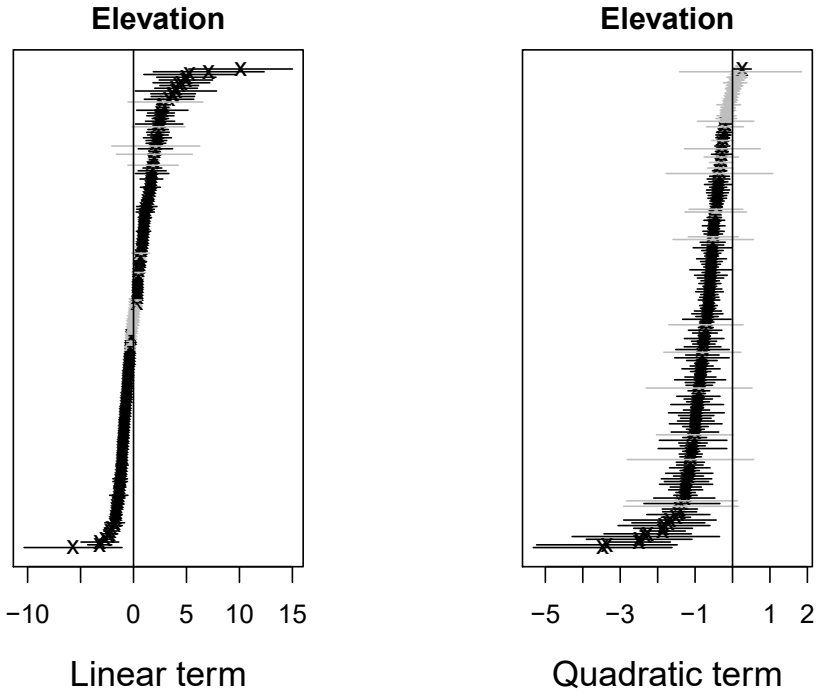


Figure S1: Plot of coefficients for elevation, from the Swiss alpine plants dataset, including 95% confidence intervals. Though no sign constraint was added for the coefficients of the quadratic term, most species exhibited concave response curves, as indicated by the negative coefficient in the second panel.

References

P. K. Dunn and G. K. Smyth. Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244, 1996. ISSN 1061-8600. doi: 10.2307/1390802.

F. K. C. Hui, D. I. Warton, J. T. Ormerod, V. Haapaniemi, and S. Taskinen. Variational Approximations for Generalized Linear Latent Variable Models. *Journal of Computational and Graphical Statistics*, 26(1): 35–43, Jan. 2017. ISSN 1061-8600. doi: 10.1080/10618600.2016.1164708.

K. Kristensen, A. Nielsen, C. W. Berg, H. Skaug, and B. Bell. TMB: Automatic Differentiation and Laplace Approximation. *Journal of Statistical Software*, 70(5), 2016. ISSN 1548-7660. doi: 10.18637/jss.v070.i05.

J. Niku, W. Brooks, R. Herliansyah, F. K. C. Hui, S. Taskinen, and D. I. Warton. Efficient estimation of generalized linear latent variable models. *PLOS ONE*, 14(5):e0216129, May 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0216129.

J. T. Ormerod and M. P. Wand. Explaining Variational Approximations. *The American Statistician*, 64(2): 140–153, May 2010. ISSN 0003-1305. doi: 10.1198/tast.2010.09058.

R. Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. ISSN 2517-6161. doi: 10.1111/j.2517-6161.1996.tb02080.x.

T. W. Yee. A New Technique for Maximum-Likelihood Canonical Gaussian Ordination. *Ecological Monographs*, 74(4):685–701, 2004. ISSN 1557-7015. doi: 10.1890/03-0078.

1.2 Model-based ordination with constrained latent variables

Model-based ordination with constrained latent variables

Bert van der Veen[123] Francis K.C. Hui[4] Knut A. Hovstad[53]
Robert B. O'Hara[23]

[1]Department of Landscape and Biodiversity, Norwegian Institute of Bioeconomy research, Trondheim, Norway
[2]Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway
[3]Centre of Biodiversity Dynamics, Norwegian University of Science and Technology, Trondheim, Norway
[4]Research School of Finance, Actuarial Studies and Statistics, The Australian National University, Canberra, Australia
[5]The Norwegian Biodiversity Information Centre, Trondheim, Norway

Summary

1. In community ecology, unconstrained ordination can be used to predict latent variables from a multivariate dataset, which generated the observed species composition. 2. Latent variables can be understood as ecological gradients, which are represented as a function of measured predictors in constrained ordination, so that ecologists can better relate species composition to the environment while reducing dimensionality of the predictors and the response data. 3. However, existing constrained ordination methods do not explicitly account for information provided by species responses, so that they have the potential to misrepresent community structure if not all predictors are measured. 4. We propose a new method for model-based ordination with constrained latent variables in the Generalized Linear Latent Variable Model framework, which incorporates both measured predictors and residual covariation to optimally represent ecological gradients. Simulations of unconstrained and constrained ordination show that the proposed method outperforms CCA and RDA. **keywords**: model-based constrained ordination, unimodal response, R-squared, joint species distribution model, reduced rank regression.

Introduction

Unconstrained ordination methods help ecologists to analyse multivariate data of species communities when measurements of the environment are missing. In ordination, species and sites are arranged by their (dis)similarity, so that in unconstrained ordination similarity in environmental conditions at sites can be inferred from species composition. For example, when species preferring wet or dry circumstances are placed at opposite sides of an ordination axis, this axis will often be interpreted to represent a gradient in soil moisture. This approach of inferring the environment of species relationships can be used to generate new hypotheses (Økland 1996). However, unconstrained ordination by design does not facilitate more exact inference of species relationships and environmental conditions at sites.

When environmental conditions are measured, e.g. such as soil moisture or mean temperature, multivariate Generalized Linear Models (MGLM, Wang *et al.* 2012) can be used to provide a more thorough understanding of species-environment relationships. However, as multivariate regression methods relate the response of each species to the predictors, the number of parameters increases rapidly with the number of species and with the number of predictors. In such instances, constrained ordination (also referred to as direct gradient analysis, ter Braak & Prentice 1988) has often been used to analyse community composition data instead. Constrained ordination assumes that an underlying complex ecological gradient can be represented as a linear combination of measured predictor variables, so that the number of parameters related to the predictors scales with the number of complex ecological gradients, and not with the number of species. Constrained ordination describes a class of methods, with two notable ones being Canonical Correspondence Analysis (CCA, ter Braak 1986) and Redundancy Analysis (RDA, Rao 1964), which (also) allow researchers to arrange sites and species by their (dis)similarity. The number of ecological gradients is often considerably less than the number of predictor variables and species (Halvorsen 2012), so that constrained ordination leads to a more feasible and potentially more insightful approach for the analysis of datasets on ecological communities with a large number of predictors and species.

The practical appeal of constrained ordination is immediately apparent in the analysis of species distributions, where bioclimatic predictor variables are often used to represent a species niche (Booth 2018). In such cases, and especially when the response data is sparse, constrained ordination can be used to reduce the number of parameters relative to standard multivariate regression (Yee & Hastie 2003). Since every added predictor variable provides more flexibility in defining the ecological gradient in constrained ordination, with a large number of predictor variables, constrained and unconstrained ordination coincide in their arrangement of sites and species (Jongman *et al.* 1995; McCune 1997; ter Braak & Šmilauer 2015).

Thus, both unconstrained and constrained ordination have their roles in the analysis of ecological com-

munities. All variation in a community can be explored with unconstrained ordination, whereas the variation due to the predictors can be explored with constrained ordination (Økland 1996; ter Braak & Šmilauer 2015). On the other hand, in situations where only a few (relevant) predictor variables are measured, i.e. some important predictors remain unmeasured, constrained ordination has the potential to misrepresent community structure as any variation not explained by the measured predictors is not accounted for in the method (Økland 1996). In turn, this motivates an approach which incorporates both: 1) modelling species responses in a reduced rank form as existing constrained ordination approaches do, and 2) a means of accounting for residual variation not accounted for by the measured predictors, as standard unconstrained ordination approaches do.

Various model-based alternatives to classical constrained ordination methods have been developed in recent years, such as those made available in the `R`-package `VGAM` (Yee & Hastie 2003; Yee 2014), and those in the `R`-package `RCIM` (Hawinkel *et al.* 2019). One of the most well-known is Reduced Rank Regression (RRR, Anderson 1951), which is a model-based approach to constrained ordination that allows users to handle a range of discrete data types, and incorporate both the linear and quadratic responses in the model (Yee 2004). However, similar to the classical constrained ordination methods CCA and RDA, RRR is a purely fixed-effects model that allows for incorporating a residual error through the specification of a response distribution, but not for error and/or residual covariation between species that is associated to an ecological gradient. This means that RRR requires the assumption that the ecological gradient can be perfectly represented by predictor variables. However, in practice it can often be unclear which predictors make up an ecological gradient, so that important predictors may remain unmeasured. As such, there is great potential for residual variation, invalidating the assumption of a perfect fit for the ecological gradient.

In this article we propose a new method for model-based ordination with constrained latent variables, which we believe has the potential to fully utilize the information provided by the measured predictors variables and species responses. In the model-based approach for constrained ordination propose here, a latent variable can be understood as a complex ecological gradient, consisting of both measured and unmeasured components, in contrast to unconstrained ordination where the latent variable always not measured. As such, the proposed model simplifies to an unconstrained ordination when no predictors are measured, or to RRR when there is no residual information left to account for after including predictor variables. The proposed approach builds on the existing framework of Generalized Linear Latent Variable models (GLLVMs, Warton *et al.* 2015), which have seen various developments in unconstrained ordination during recent years (Hui *et al.* 2015; Hui 2016, 2017; Niku *et al.* 2019; Hoegh & Roberts 2020; Damgaard *et al.* 2020; van der Veen *et al.* 2021; Zeng *et al.* 2021). However, GLLVMs still lack an implementation when it comes to applications in constrained ordination. Here, we extend the GLLVM framework for model-based ordination

to the constrained case. Performing constrained ordination in the GLLVM framework allows us to relax the assumption of a perfect fit of predictors to the ecological gradient, so that the latent variables are both a function of the predictors as in the constrained case, and include residual variation provided by the response data as in the unconstrained case. In doing so, this approach allows the latent variables to better represent ecological gradients.

Through a series of simulations based on multivariate normal, presence-absence and count data, we demonstrate that even in the presence of many predictor variables, estimating species responses using RRR can perform just as well if not better than in multivariate regression (e.g., Wang *et al.* 2012) while using fewer parameters. These simulations provide a basis for the evaluation of the dimensionality of community structure and species responses. We additionally compare our proposed GLLVM approach with two popular constrained ordination methods, CCA and RDA, assessing their capability to retrieve the true ecological gradients and species responses in the presence and absence of residual variation and fixed-effects. We show that in the presence and absence of residual variation the proposed GLLVM with constrained latent variables performs similar to, if not better than, CCA and RDA in retrieving the ecological gradients and species responses. Further, when the predictors are unrelated to the ecological gradients, e.g., when the wrong predictors have been measured, the proposes GLLVM with constrained latent variables outperforms CCA and RDA.

Finally, we use two real datasets from species communities to demonstrate use of the proposed GLLVM with constrained latent variables: a dataset of alpine plants on an elevation gradient in Switzerland (D'Amen *et al.* 2017), and a dataset of vascular plants in semi-natural grasslands collected in Norway. An easy-to-use software implementation for model-based ordination with constrained latent variables is available on CRAN in the `gllvm` R-package.

Model-formulation

For a multivariate dataset y_{ij} consisting of observations recorded for species $j = 1 \dots p$ and sites $i = 1 \dots n$, the proposed GLLVM with constrained latent variables is defined by the following mean model. Let $g(\cdot)$ generically denote a link function that connects the mean of the assumed response distribution (e.g., the Bernoulli distribution for presence-absence data or the negative binomial for overdispersed counts) to a linear predictor η_{ij}, with a vector $\boldsymbol{x}_{lv,i}$ of $k = 1 \dots K$ measured predictor variables e.g., solar radiation or available cover. Then we formulate the model with $q = 1 \dots d$ constrained latent variables as:

$$g\{\mathrm{E}(y_{ij}|\boldsymbol{x}_{lv,i}, \boldsymbol{\epsilon}_i)\} = \beta_{0j} + \boldsymbol{x}_{lv,i}^\top \boldsymbol{B}\boldsymbol{\gamma}_j + \boldsymbol{\epsilon}_i^\top \boldsymbol{\gamma}_j, \tag{1}$$

where β_{0j} is an intercept for each species j, and $\boldsymbol{B}$ is a $K \times d$ matrix of slopes per predictor and latent variable. A constrained latent variable can be understood as a complex ecological gradient, of which some components are (un)measured. As such, the vector $\boldsymbol{\gamma}_j$ includes relative species responses to both the measured component of the constrained latent variables $\boldsymbol{B}^\top \boldsymbol{x}_{lv,i}$ and their residual variation $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\sigma}^2 \mathrm{I}_d)$, where I_d is a $d \times d$ identity matrix. Note, the predictors $\boldsymbol{x}_{lv,i}$ exclude an intercept term, for reasons of parameter identifiability, due to the presence of species-specific intercepts β_{0j}. More importantly, the model in equation (1) can instead be formulated in terms of a latent variable $\boldsymbol{z}_i = \boldsymbol{B}^\top \boldsymbol{x}_{lv,i} + \boldsymbol{\epsilon}_i$, similar to the model form in unconstrained ordination methods (see e.g., Hui *et al.* 2015; van der Veen *et al.* 2021):

$$g\{\mathrm{E}(y_{ij}|\boldsymbol{z}_i)\} = \beta_{0j} + \boldsymbol{z}_i^\top \boldsymbol{\gamma}_j, \tag{2}$$

where $\boldsymbol{z}_i \sim \mathcal{N}(\boldsymbol{B}^\top \boldsymbol{x}_{lv,i}, \boldsymbol{\sigma}^2 \mathrm{I}_d)$. The models in equation (1) and equation (2) can be straightforwardly extended with row-intercepts to model community composition instead, though we have chosen to omit that term here for ease of presentation.

Without the residual term for the latent variables, equation (1) is an ordinary reduced rank regression (RRR, Anderson 1951), similar to RR-VGLMs implemented in the `R`-package `VGAM` (Yee & Hastie 2003; Yee 2004), or classical constrained ordination methods (ter Braak & Šmilauer 2015). Then, the latent variables are represented only by the fixed-effects term $\boldsymbol{B}^\top \boldsymbol{x}_{lv,i}$ which represents a constrained ordination axis, and serves to reduce the dimensionality of the number of predictors K. Compared to standard multivariate regression, RRR can serve to reduce the number of parameters, as the number of parameters can be especially difficult to estimate for large K and small n. The matrix of coefficients for species $j = 1 \dots p$ from a multivariate regression $\boldsymbol{\beta}_j$ can be reconstructed from an RRR as $\boldsymbol{\beta}_j = \boldsymbol{B}\boldsymbol{\gamma}_j$ with accompanying standard errors (see Appendix S1). The number of parameters in the model is then $p + d(p + K) - (d + d^2)/2$ for rank d, which can often be a more realistic assumption for ecological community data that tend to be sparse on information. Note that in some cases, the number of parameters in the reduced rank model can exceed the number of parameters used to model species responses in full rank (e.g., when $d = K$), though most commonly we assume $d < K \ll p$ so that rank-reduction is achieved.

When measurements of the environment are missing entirely, or when the predictors are unrelated to the ordination and have slopes close to zero, in essence when $\boldsymbol{B}^\top \boldsymbol{x}_{lv,i} = \boldsymbol{0}$, then the model in equation (1) simplifies to an unconstrained ordination. In the method for model-based ordination with constrained latent variables proposed here, the term $\boldsymbol{\epsilon}_i$ is used to account for the discrepancy between the true latent variable, and the latent variable that can be predicted using the predictor variables alone. In summary, the model proposed here performs simultaneous constrained and unconstrained ordination when predictor variables are

included.

The model in equation (1) can be extended to include additional (separate) predictors, resulting in a partial constrained ordination similar to ter Braak (1988):

$$g\{\mathrm{E}(y_{ij}|\boldsymbol{x}_i, \boldsymbol{x}_{lv,i}, \boldsymbol{\epsilon}_i)\} = \beta_{0j} + \boldsymbol{x}_i^\top \boldsymbol{\kappa}_j + \boldsymbol{x}_{lv,i}^\top \boldsymbol{B}\boldsymbol{\gamma}_j + \boldsymbol{\epsilon}_i^\top \boldsymbol{\gamma}_j, \tag{3}$$

where $\boldsymbol{\kappa}_j$ are species coefficients for the predictors $\boldsymbol{x}_i$, and where we additionally assume that $\boldsymbol{x}_i$ and $\boldsymbol{x}_{lv,i}$ do not include the same predictor variables for reasons of parameter identifiability. Here, the effect of $\boldsymbol{x}_i$ is excluded from the constrained ordination, and are included so that the resulting ordination is interpreted conditionally on the predictors $\boldsymbol{x}_i$ and species slopes $\boldsymbol{\kappa}_j$.

The models presented so far assume that species respond linearly to the latent variables. However, it is widely acknowledged that species respond to the environment unimodally (see e.g., ter Braak 1987). van der Veen *et al.* (2021) recently presented a method for model-based ordination with quadratic responses, as a means to model species-specific environmental tolerances. All models presented here can be extended in a similar fashion, e.g. with quadratic response to the ecological gradients, which we further elaborate on in Appendix S2. Furthermore, classical constrained ordination methods are infamous for their increased variance of the $\boldsymbol{B}$ parameters. Although that might be due to the lack of a maximum likelihood solution, the model here can be extended using a random slope formulation in order to regularize the predictor slopes, and retrieve parameter estimates with reduced variance (see Appendix S3).

The model in equation (1) is unidentifiable without additional constraints, due to the freely varying scale parameters $\boldsymbol{\sigma}$ for the latent variables. Consider a matrix $\boldsymbol{\Gamma}$ that includes all species loadings $\boldsymbol{\gamma}_j$ as row vectors, for which we fix the upper triangular entries to zero for reasons of parameter identifiability, as is usual for GLLVMs (Hui *et al.* 2015). In standard formulation of GLLVMs, due to scale invariance, the latent variables are assumed to have unit variance. Then, the species slopes $\boldsymbol{\gamma}_j$ additionally serve to determine the scale of the ordination. However, in model-based ordination with constrained latent variables, the species loadings are shared for two terms, so that without extra constraints they regulate the relative scale of the second and third term in equation (1). In cases where either the fixed-effects term or the residual term is zero, this requires the model to compensate by increasing the magnitude of the species loadings, and e.g., decreasing the magnitude of $\boldsymbol{B}$ if the true fixed-effects term is non-zero but the residual term is not. Therefore, we additionally choose to fix one parameter per latent variable to facilitate including freely varying scale parameters for the latent variables. Here, we choose the diagonal entries of $\boldsymbol{\Gamma}$, such that in $\boldsymbol{\Gamma}$ there are only $(p-d)d+d(d-1)/2$ parameters to estimate. This choice of the diagonal elements is arbitrary, and different elements could be chosen instead. However, the current choice is guided by the magnitude of

the different parameters in the model, as now $\boldsymbol{B}$ determines the scale of the first (fixed-effects) term, so that it is (close to) zero when the predictors have no effect on the ordination. Similarly, the vector of residual standard deviations $\boldsymbol{\sigma}$ then determines the scale of the residual term, so that it is zero when there is no residual necessary in the ordination (i.e. when the predictors perfectly represent the latent variable, as in RRR).

Consequently, the vector of standard deviations for the residual of the latent variable $\boldsymbol{\sigma}$ can additionally be used to determine when latent variables are nearly redundant, or additionally for a measure of (residual) gradient length, or to develop a method of regularization in GLLVMs. In comparison, van der Veen *et al.* (2021) considered the scale of the latent variables relative to the median tolerance of species curves being one. However, this has no meaning in models with the linear responses, and is difficult to implement in practice. Finally, it is important to note that this choice of identifiability constraint does not diminish the overall flexibility of the model, but merely clarifies the interpretation of the parameters: in essence, the latent variables are stretched or contracted so that certain species loadings equal one.

Parameter estimation and model fitting

Since the proposed method is a type of GLLVM, we are required to choose an appropriate distribution to model the species observations (and their associated mean-variance relationship, see e.g., Warton & Hui 2017). For example, a Poisson or negative binomial distribution with log-link function for counts, a Binomial distribution with probit link-function for presence-absence data, or alternatively a Tweedie distribution with log-link function for biomass data. Similar to other GLLVMs proposed for model-based ordination in the literature, the residual error terms $\boldsymbol{\epsilon}_i$ are assumed to be normally distributed random variables, which thus needs to be integrated over. Consequently, the marginal log-likelihood of the proposed GLLVM with constrained latent variables as in (1) is written as:

$$\mathcal{L}(\Theta) = \sum_{i=1}^{n} \log\left\{ \int \prod_{j=1}^{p} f\Big(y_{ij} | \boldsymbol{x}_i, \boldsymbol{\epsilon}_i, \Theta \Big) h\Big(\boldsymbol{\epsilon}_i \Big) d\boldsymbol{\epsilon}_i \right\}, \tag{4}$$

where $f(y_{ij}|\boldsymbol{x}_{lv,i}, \boldsymbol{\epsilon}_i, \Theta)$ is the distribution of the responses conditional on the predictors $\boldsymbol{x}_{lv,i}$, the constrained residual error term $\boldsymbol{\epsilon}_i$, and a vector Θ. The vector Θ includes all parameters and an estimate of $\mathrm{h}(\boldsymbol{\epsilon}_i|y_{ij})$ with related variational parameters if applicable. The residual error terms are assumed to follow a multivariate normal distribution $h(\boldsymbol{\epsilon}_i) = \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}^2 \mathrm{I}_d)$. The integration can be performed with methods previously developed for estimation and inference in GLLVMs, such as the Laplace approximation (Niku *et al.* 2017) or Variational Approximations (VA, Hui *et al.* 2017; van der Veen *et al.* 2021) and using Template Model

Builder (Kristensen *et al.* 2016). Below, the models in the simulation studies and examples are fitted using VA.

Initial values

Both with and without residual error term, the algorithm used to fit the models presented here is sensitive to the initial values. In this article, we adapt the approach used in the `gllvm` R-package to overcome this, and obtain reasonable starting values. Specifically, for ordinary RRR, we followed a similar procedure to that described by Files *et al.* (2019), where we generate starting values for $\boldsymbol{B}$ and $\boldsymbol{\gamma}_j$ by first fitting a multivariate linear model with predictor variables to the Dunn-Smyth residuals (Dunn & Smyth 1996) of an intercept-only MGLM. We then performed a QR-decomposition on the matrix of regression coefficients to obtain the starting values for $\boldsymbol{B}$ and $\boldsymbol{\gamma}_j$. For constrained latent variables, we performed a factor analysis on the Dunn-Smyth residuals of an intercept-only MGLM, and then regress the estimated factor scores to receive initial values for the predictor slopes $\boldsymbol{B}$. The residuals of the regression provided initial values for $\boldsymbol{\epsilon}_i$, and the loadings from the factor analysis were taken as the initial values for $\boldsymbol{\gamma}_j$.

Inference

In this section we present various tools for inference and prediction for the proposed method of model-based ordination with constrained latent variables.

Constrained ordination diagram

Separate ordination diagrams can be constructed for both terms in equation (1) to explore species relationships for the predictors and for residual variation, or an ordination diagram can be constructed including both terms to present species co-occurrence patterns within a single plot. In GLLVMs with unconstrained latent variables, we can obtain predictions of the residuals $\boldsymbol{\epsilon}_i$ using e.g., the means of variational distributions (Hui *et al.* 2017) or the maximum a-posteriori prediction from the Laplace approximation (Niku *et al.* 2017). Note however these latent variables are assumed to fully consist of residual information as discussed previously. In contrast, with model-based ordination using constrained latent variables, we instead consider using the predicted site scores $\boldsymbol{z}_i$ for ordination, which can be constructed based on the predicted $\boldsymbol{\epsilon}_i$'s along with the estimated value for $\boldsymbol{B}$. We consider these site scores similar to the weighted average (WA) scores (Palmer 1993; McCune 1997) provided by classical constrained ordination methods such as CCA or RDA, since the residual term $\boldsymbol{\epsilon}_i$, accounts for variation in the response not explained by the predictor variables. As

such, $\boldsymbol{B}^\top \boldsymbol{x}_{lv,i}$ are linear predictor (LC) scores, which we will refer to here as "marginal" scores, that do not include additional information on the latent variable provided by the response data (Palmer 1993). Similarly, we denote site scores for the residual term as "residual" scores and site scores that include both terms as "conditional" scores. Note that RRR always only includes LC scores and unconstrained ordination always includes residual scores. LC scores are not generally recommended for inference by community ecologists for classical constrained ordination methods (McCune 1997).

A constrained ordination diagram with conditional site scores will, in many instances, provide a similar ordination as when latent variables are assumed to be unconstrained. On the other hand, constrained ordination allows the predictor effects to be represented in an ordination diagram, in the form of arrows based on the rows of $\boldsymbol{B}$. The length of the arrow is proportional to the magnitude of the parameter estimate, so that the predictor with the largest estimate is presented as the longest arrow, although note that we correct the arrow length using the standard deviation of each predictor (e.g. as in Figure 3 below). Statistical uncertainty of the slope estimates for the predictors can be further represented using the colour of the arrows, for example by colouring the arrow less intensely for predictor slope estimates for which the corresponding confidence interval includes zero for at least one of the ordination dimensions.

In an ordination diagram, the predicted site scores are plotted to represent (dis)similarity between sites in an ordination. Furthermore, Niku (2020) constructed prediction regions using the Conditional Mean Squared Error of Predictions (CMSEPs, Booth & Hobert 1998) to represent the statistical uncertainty of the site scores in an ordination diagram. To fully and properly convey confidence in the dissimilarity of sites, we adopt the same approach, but adapt the calculation for the case of constrained ordination (see Appendix S4 for details of the calculation). These prediction intervals can be used to provide a larger degree of certainty in the dissimilarity of sites, and tend to be larger than in unconstrained ordination, as they represent both the uncertainty of the fixed-effects and of the residual error term.

Model selection

As the number of predictors increases, and flexibility is added in the modelling of the site scores and species responses, the standard deviations of the residual term $\boldsymbol{\sigma}$ are likely to get smaller. Determining the optimal number of latent variables and the most suitable predictor variables for a constrained ordination is thus an important problem for our method, although it can be a challenging exercise as the number of potential models may be quite large. In the `vegan` `R`-package various tools are available to find the combination of predictor variables that optimally represents the latent variable, such as stepwise selection using permutation P-values or an adjusted $R^2_{\boldsymbol{B}}$ (Oksanen *et al.* 2020). As a model-based approach, we can leverage conventional

methods such as hypothesis testing, information criteria (Burnham & Anderson 2002), residual diagnostics (Hartig 2021) among others for assessing the optimal number of latent variables and predictors, predictions, as well as assessing other model assumptions such as the distribution of the responses. For example, the importance of predictors in a model-based ordination with constrained latent variables can be assessed with use of a Wald-statistic and associated p-values, or with confidence intervals. We illustrate an example of determining predictor importance later on in our applications of two real datasets.

Similarly, the question of whether to perform constrained or unconstrained ordination, and thus also which type of site scores is more suitable for representing the ecological community, can be solved using model-selection tools such as information criteria.

Predictor importance

Similarly to other GLLVMs, the residual covariance matrix associated with the latent variables can be calculated (see Appendix S5) to examine species associations and determine the residual variation in the response, beyond that due to the measured predictors. By fitting a second unconstrained model, the variation explained by the predictors in the response can also be determined, similar to the approach presented by Warton *et al.* (2015) based on relative differences in the trace of the residual covariance matrix. Here however, with model-based ordination using constrained latent variables, we focus on determining the importance of predictors in explaining the latent variables. Since the latent variables are by definition unmeasured, calculating importance of the predictors through e.g. a partial $R^2_{\boldsymbol{B}}$ as in ordinary linear regression, is not directly possible. As such, to assess the importance of predictors in explaining the latent variables, we adopt an approach similar to that presented by Edwards *et al.* (2008), which also avoids having to fit a second model (which can be computationally intensive for a large number of sites n and species p). Specifically, Edwards *et al.* (2008) developed a measure of $R^2_{\boldsymbol{B}}$ for linear mixed-effects models based on the fit of a single model, which Jaeger *et al.* (2017) extended to the generalized linear mixed-effects model and implemented in the `r2glmm` `R`-package, using a multivariate Wald-statistic for the testing of fixed-effects. Thus, the proportion of (generalized) variance explained by all predictors for all latent variables is:

$$R^2_{\boldsymbol{B}} = \frac{\omega \text{vec}(\hat{\boldsymbol{B}})^\top (\hat{\boldsymbol{\Sigma}})^{-1} \text{vec}(\hat{\boldsymbol{B}})(dK)^{-1}}{1 + \omega \text{vec}(\hat{\boldsymbol{B}})^\top (\hat{\boldsymbol{\Sigma}})^{-1} \text{vec}(\hat{\boldsymbol{B}})(dK)^{-1}}, \tag{5}$$

where $\text{vec}(\hat{\boldsymbol{B}})$ is a vectorized version of the matrix of estimated predictor slopes $\hat{\boldsymbol{B}}$ with corresponding estimated covariance matrix $\hat{\boldsymbol{\Sigma}}$, and ω is a ratio based on the residual degrees of freedom: $\frac{dK}{dn-dK}$, for the total number of sites n, the number of constrained latent variables d, and the number of predictors K. Note this $R^2_{\boldsymbol{B}}$ can also be calculated on a per predictor variable basis (with numerator degrees of freedom d), or

per latent variable and predictor (with unit numerator degrees of freedom), to retrieve a semi-partial $R^2_{\boldsymbol{B}}$. This semi-partial $R^2_{\boldsymbol{B}}$ is interpreted as the capability of that predictor to explain the (generalized) residual variation unaccounted for after accounting for all other predictors in the full model (Edwards *et al.* 2008).

To summarize, a high semi-partial $R^2_{\boldsymbol{B}}$ indicates importance of a predictor in explaining the latent variables. We demonstrate use of the semi-partial $R^2_{\boldsymbol{B}}$ in the real data examples below.

Simulation studies

We performed two separate simulation studies for the proposed GLLVM with constrained latent variables: 1) we simulated from a MGLM (i.e. with full rank species responses) using bioclimatic predictor variables, and compared the accuracy of RRR (i.e. model-based ordination with constrained latent variables and $\boldsymbol{\epsilon}_i = \mathbf{0}$) and a MGLM to retrieve the true species responses, and 2) we simulated unconstrained and constrained ordinations to compare the capability of the proposed GLLVM to retrieve the true latent variables $\boldsymbol{z}_i = \boldsymbol{B}^\top \boldsymbol{x}_{lv,i} + \boldsymbol{\epsilon}_i$ and species loadings $\boldsymbol{\gamma}_j$, in comparison to the WA scores from CCA and RDA. Since it is more difficult to accurately predict the latent variables when the number of species is small, and since more sites provide more information to estimate species responses, in both simulations we included few species but more sites. `R`-code for the simulation studies is included in Appendix S6.

In our first simulation study, to get a realistic collinearity structure in the predictor variables, we simulated 1000 random points across the European Union, Switzerland and Norway, using the `sp` `R`-package (Pebesma & Bivand 2005), at which we retrieved 19 bioclimatic variables using the `raster` `R`-package (Hijmans 2020). Afterwards, we simulated $K = 19$ bioclimatic predictor variables at $n = 100$ new sites, from a multivariate normal distribution with a zero mean vector and the covariance matrix set equal to the sample covariance matrix of the 1000 random points. We then standardized the predictor variables to have mean zero and variance one. We removed one predictor that was almost fully collinear with another predictor, as it resulted in numerical issues, so that the final number of predictors was $K = 18$. We then used the extended hunting spider dataset with $n = 100$ sites and $p = 12$ species from van der Aart & Smeek-Enserink (1975), to fit a MGLM with a Poisson distribution and log-link for the count responses and including the 18 predictor variables above, and subsequently used the estimated species-specific slopes from this model as the true slopes in our simulation. Additionally, we simulated species-specific intercepts from the uniform distribution Uniform$(-1, 1)$ corresponding to species of low abundance or occurrence. We simulated 1000 datasets assuming either Gaussian responses, Poisson counts or Bernoulli presence-absence responses. The variance associated with Gaussian responses was assumed to be one. For each simulated dataset, we fitted a MGLM along with RRR with 2-8 latent variables in equation (1). The rank of the matrix of species responses

can maximally be $\min(p, K)$, so that it was 12 here. The MGLMs included the same number of parameters as the true model, namely 228 parameters, whereas the fitted RRR models included 69 parameters for $d = 2$, 96 for $d = 3$, 122 for $d = 4$, and 147 for $d = 5$, 171 for $d = 6$, 194 for $d = 7$, and 216 for $d = 8$. We did not include a rank 9 model in our example, as it would have exceeded the number of parameters in the MGLM (237 parameters versus 228 parameters). Finally, we calculated the symmetric Procrustes error between the true and and estimated matrix of species responses for each simulated dataset, as to compare accuracy of the methods in retrieving species responses, using the `vegan` `R`-package (Oksanen *et al.* 2020).

The results of the first simulation study are summarized in Figure 1, which show that as the number of latent variables in the RRR increased, the accuracy of the estimated species responses improved. At rank 5, the accuracy was similar to that of the MGLM, yet the number of parameters included in the model was about 65% of that of the true model. In general, the estimated species responses were accurately estimated across all distributions, though for Bernoulli responses the error was generally higher than for Gaussian or Poisson responses.

For the second simulation study, we considered datasets with $n = 100$ sites and with $p = 30$ species. We simulated three forms of the model in equation (1): with 1) non-zero predictor slopes $\boldsymbol{B}$ and residuals $\boldsymbol{\epsilon}_i$, 2) non-zero predictor slopes without the residuals, i.e. RRR, and 3) with $\boldsymbol{B} = \boldsymbol{0}$ and non-zero residuals, i.e. an unconstrained ordination.

To construct the true model, we first simulated $K = 5$ predictor variables following a multivariate standard normal distribution. Next, we generated the true slope coefficients $\boldsymbol{B}$ by applying a factor analysis to the simulated predictor variables, with two dimensions. We simulated the true intercepts β_{0j} from Uniform$(-1, 1)$, and species coefficients for the latent variables $\boldsymbol{\gamma}_j$ independently from Uniform$(-2, 2)$. Finally, we simulated the constrained residual error $\boldsymbol{\epsilon}_i$ by first sampling from a bivariate standard normal distribution, after which we regressed the sampled realization against the simulated predictor variables, and used the residual from the regression as the residual error in the true model. This ensures that the true residual error $\boldsymbol{\epsilon}_i$ was independent of the simulated predictor variables by construction. We simulated 1000 datasets each from the Gaussian, Bernoulli, and Poisson distributions. The variance associated with Gaussian responses was again assumed to be one. To each dataset, we fitted a GLLVM with two constrained latent variables while CCA was fitted to the datasets with Bernoulli and Poisson responses, and RDA to the datasets with Gaussian distributed responses. Both classical ordination methods were fitted using the `vegan` `R`-package, which we also used for the calculation of a symmetric Procrustes error between the simulated latent variables and the latent variables retrieved from the proposed GLLVM, CCA, and RDA, and the same for the species loadings (Peres-Neto & Jackson 2001; Oksanen *et al.* 2020). The results are summarized in Figure 2.

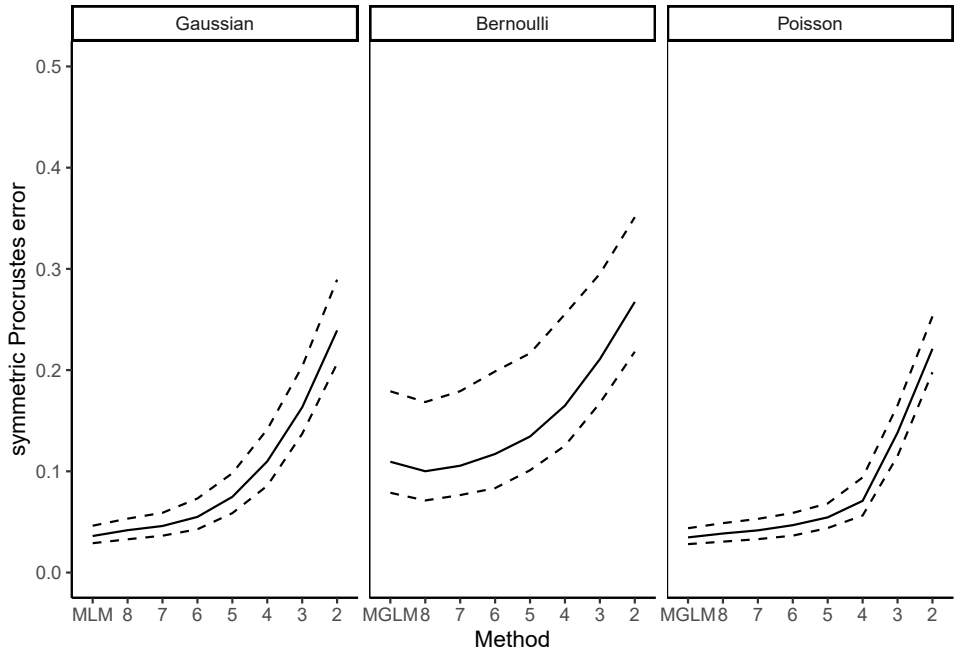


Figure 1: Simulation results for the 1000 multivariate GLMs and various reduced rank regressions fitted to Gaussian, Bernoulli and Poisson response datasets simulated from a multivariate GLM, with $n = 100$ sites and $p = 12$ species. The true model contained 18 bioclimatic predictor variables with a realistic degree of collinearity. The symmetric Procrustes error of the estimated and true species responses is shown on the y-axis. Numbers 2-8 indicate reduced rank models with that rank. The solid line represents the median symmetric Procrustes error, and dashed lines the first and third quartiles.

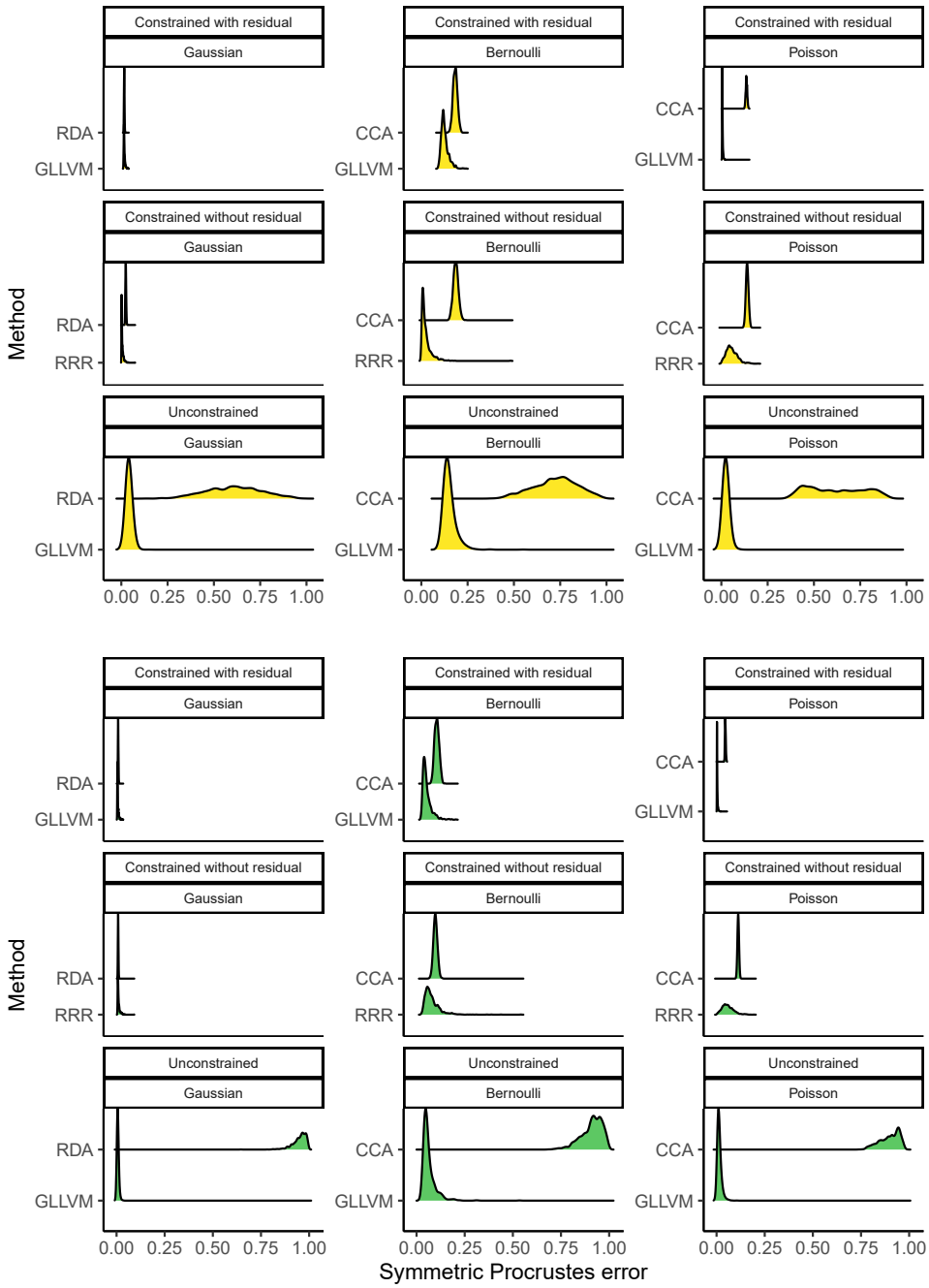


Figure 2: Results for ordination methods fitted to 1000 simulated datasets with Gaussian, Bernoulli, and Poisson response datasets. Simulations labled 'unconstrained' followed the same true model as simulations labled 'constrained', but instead slopes for the predictor variables were fixed to zero i.e., $\boldsymbol{B} = \boldsymbol{0}$. For simulations labled 'constrained without residual', the true model was the same as that of 'constrained', but without the residual variation i.e., $\boldsymbol{\epsilon}_i = \boldsymbol{0}$. The procrustes error of the latent variables $\boldsymbol{z}_i$ (yellow) and species loadings $\boldsymbol{\gamma}_j$ are presented.

In general, for the proposed GLLVM with constrained latent variables, with and without residual (i.e. RRR), but also without fixed-effects (i.e. unconstrained ordination), we consistently and with little variability managed to retrieve the true latent variables $\boldsymbol{z}_i$ and species loadings $\boldsymbol{\gamma}_j$. For constrained ordination and for Gaussian distributed responses, RDA compared to GLLVMs and RRR performed similarly. However, when RDA was fitted to datasets where the predictor variables had no relation to the latent variables, it was unable to retrieve the latent variables or species loadings. In all cases, GLLVMs and RRR performed better than CCA. Similarly to RDA, CCA was not able to retrieve the latent variables or the species loadings if the predictor variables had no relation to the latent variables.

Worked examples

We demonstrate applications for the proposed GLLVM with constrained latent variables on two ecological datasets: 1) a dataset of Swiss alpine plants (D'Amen *et al.* 2018), and 2) a dataset of vascular plants collected in Levanger, Norway.

Swiss alpine plants

The first example includes a presence-absence dataset of alpine plants, collected in the western Swiss Alps. The dataset was collected on a strong elevation gradient, including sites in both lowland and alpine environments (D'Amen *et al.* 2018). In total the dataset includes $n = 791$ plots and $p = 175$ plant species, after excluding plots with fewer than two observations and species with less than three presence observations (D'Amen *et al.* 2018). Six predictor variables were included in the study: degree days above zero, slope, moisture index, total solar radiation over the year, topography, and elevation. All predictors were standardized to have mean zero and variance one.

Fitting a range of models while testing for the optimal number of dimensions and predictors would be computationally burdensome and time consuming, so for demonstration purposes we here fit a model with $d = 2$ latent variables and assuming a Bernoulli distribution for the responses, using all predictors, and with one quadratic coefficient per latent variable. Previously, van der Veen *et al.* (2021) found that using a quadratic response model lead to better predictions of the ecological gradient for this dataset, so we adopt that approach here. We then base our inference of predictor importance on the confidence interval of predictor slopes i.e., a Wald-statistic with accompanying p-values, and the approach for $R^2_{\boldsymbol{B}}$ presented above.

The results from the constrained ordination are summarized in Appendix S7: Table S1. Similar to van der Veen *et al.* (2021), elevation and degree days above zero were the two predictors most related to the two

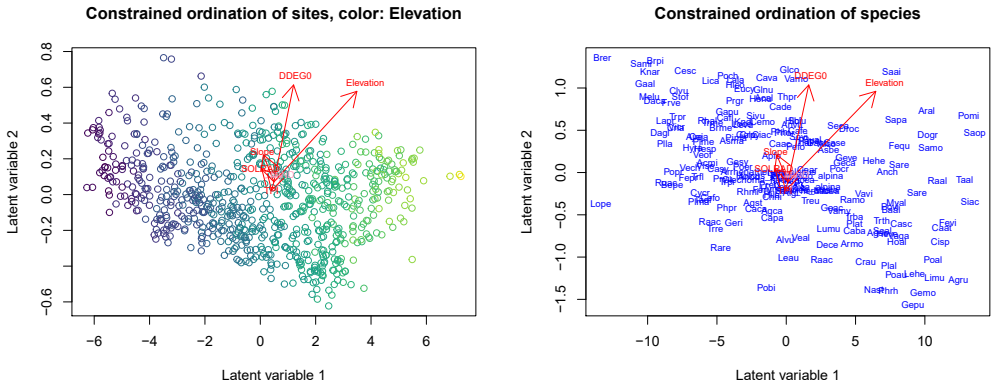


Figure 3: Constrained ordination diagram for the Swiss alpine plants data. Darker colors in the left plot plot indicate low elevation, whereas lighter colors indicate sites at higher elevation. Arrows represent predictor slopes for each latent variable, with arrow length being proportional to the magnitude of the slope estimate. Arrows shown in pink represent slope estimates of which the confidence intervals included zero for one of the dimensions. The latent variables have been rotated to principal direction, so that the first latent variable explains most variation. In the right plot, the abbreviated species names represent optima from the quadratic response model, which are drawn as arrows if they are too far removed from the latent variable. Detailed results, and a list of species names, are included in Appendix S7: Table S1, Table S2.

predicted latent variables. For the first latent variable, the slope for the elevation predictor was four times as large as the slope for degree days above zero. For the second latent variable, the magnitude of the slopes for elevation and degree days above zero was similar. The results are visually presented in Figure 3.

Based on the fitted GLLVM with constrained latent variables containing all six predictor variables, the residual standard deviations were 0.00 (95% confidence interval: -0.02, 0.02) and 0.34 (0.32, 0.36), indicating that the first latent variable could be fully represented by the predictors. The $R^2_{\boldsymbol{B}}$ for the linear regressions of the latent variables was 0.34. The semi-partial $R^2_{\boldsymbol{B}}$ for the predictors was 0.08 (degree days above zero), 0.14 (slope), 0.02 (moisture index), 0.06 (solar radiation), 0.04 (topography), and 0.15 (elevation), indicating that elevation and slope were most important for representing the latent variables. Our results overall were similar to those for CCA, where the largest correlation of a predictor with the first axis was for elevation, and for the second axis slope (see also Appendix S7: Figure S1).

Semi-natural grasslands in Norway

The second example contains observations of vascular plants collected at Levanger, Norway in 2008. In total $n = 132$ plots of $1m^2$ were pseudo-randomly positioned across five "zones". For each zone, coordinates were sampled randomly, and plants were recorded if the coordinate was located in a semi-natural grassland. We

excluded sites for which there were no soil measurements available, so that the final dataset included $n = 116$ sites. In total, the dataset included observations of $p = 132$ vascular plants, but we chose to exclude species with fewer than 3 observations, so that the final dataset included observations of $p = 64$ vascular plant species. Some of the grasslands were grazed by sheep and cattle, whereas others were abandoned and had no management. The study area was approximately five kilometres long and four kilometres wide, and located along a ridge with an east-west direction. The collected data is in the form of percentage cover per species in a plot, so that the total coverage of all plants in a plot can exceed 100 percent. In total, the dataset included 36 different predictor variables of various water-soluble soil nutrients such as pH, phosphorus, potassium, calcium, and organic matter content for the first 0 - 10 centimetre of soil and additionally for the next 10 - 20 centimetre layer. Though the soil variables for 0 - 10- and 10 - 20 centimetres soil depth to some degree are collinear, they represent different aspects of the edaphic site conditions; at 0 - 10 centimetre deep the soil properties are more affected by current conditions such as management and present vegetation. The results from soil samples at 10 - 20 centimetres depth were included to represent the mineral content of deeper soil layers less influenced by present management and vegetation. Below, soil variables for the first 0 - 10 centimetres are indicated by a one, and for the 10 - 20 centimetres layer with a two. Various other predictor variables were also recorded, such as whether a plot was grazed or abandoned, the cover of trees, shrubs, litter, and the height of the different layers, slope, and aspect. For demonstration purposes, we focus on the soil property variables, and on the effects of grazing, so that the final number of predictors in the constrained ordination was $K = 10$.

We fitted a model with two constrained latent variables, assuming a Tweedie distribution with power parameter 1.3, as that seemed to provide the best fit as determined from re-fitting and examining of residual plots. A Tweedie distribution has the potential to predict percentages larger than 100 here, and as such it might be more realistic to fit the model using a beta distribution. In contrast, the Tweedie will provides a more flexible mean-variance relationship than the beta distribution. Additionally, the data here include zeros, which are not possible to include with a beta distribution. To demonstrate the possibility of additional predictors, as per equation (3), we included whether a plot was grazed or not as an additional fixed effect. Thus, the ordination will be conditional on the effect of grazing, so that by colouring sites based on this predictor, it should then become clear that the effect has been accounted for outside of the ordination. More details from the constrained model for the Levanger grasslands dataset, including the effect of grazing on individual plant species, and a table of estimates and standard errors for the ordination, are included in Appendix S7: Table S3, Figure S3.

The residual standard deviations of the constrained latent variables were 1.11 (95% confidence interval: 0.41, 1.82) and 0.78 (0.17, 1.40). The predictors poorly explained the latent variables, as indicated by the

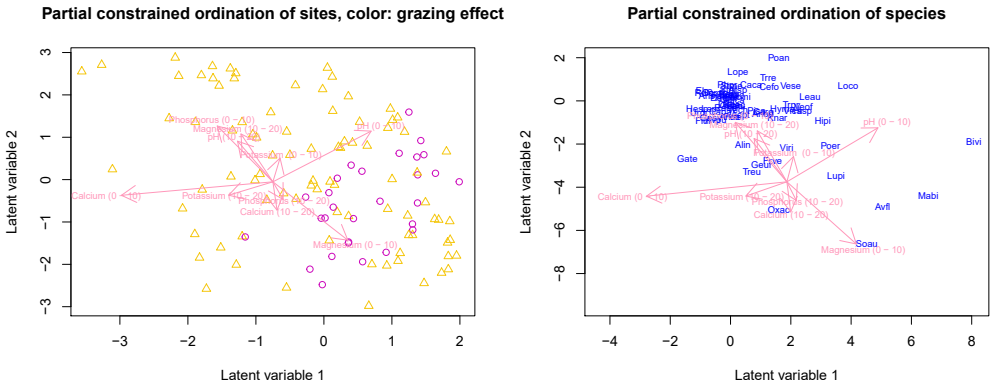


Figure 4: Constrained ordination of sites (left) and species (right) from the Levanger grasslands dataset. The constrained ordination is conditional on the effect of grazing, so that the effect excluded from the ordination diagram. To emphasize this, sites have been marked by their classification: yellow triangles are the reference plots, whereas purple circles indicate grazed plots. By conditioning on the effect of grazing, this constrained ordination now solely focuses on arranging sites (and potentially species) based on soil properties. Estimates and standard errors of the effects are provided in Appendix S7: Table S3, though none of the effects were statistically significant for both dimensions. A list of species names is included in Appendix S7: Table S4.

low $R^2_{\boldsymbol{B}}$ for the latent variables of 0.08, so that the semi-partial $R^2_{\boldsymbol{B}}$ of all predictors was close to zero, but largest for calcium. For CCA, the largest correlation for the first axis was for pH (0 - 10) and Ca (10 - 20), and K (10 - 20) for the second axis. The variation that could be explained by the predictors was less than 18% of the overall variation (see also Appendix S7: Figure S2).

In the light of the above results with $R^2_{\boldsymbol{B}}$, it was not surprising that most of the estimated predictor slopes i.e., elements of $\boldsymbol{B}$, were accompanied by a large statistical uncertainty, so that no predictor estimates were significantly different from zero for both latent variables simultaneously (see Appendix S7: Table S3). For the first dimension, the effect of pH at soil depth 0 - 10 centimetres was statistically significant, as well as the effect of calcium. For the second dimension, the effects of potassium (0 - 10 cm) and magnesium (0 - 10 cm and 10 - 20 cm) were statistically significant. The grazed and abandoned sites occupy a similar space in the ordination diagram (Figure 4), as was expected when performing a partial constrained ordination.

Discussion

In this article, we present a new method for model-based ordination with constrained latent variables, or alternatively for estimating species responses in a reduced-rank form, using the GLLVM framework (Warton *et al.* 2015). Constrained ordination allows ecologists to order sites and species using measured predictors,

and as such to better examine species-environment relationships (Ter Braak 1987). When there is no effect of the predictors, the model proposed here simplifies to an unconstrained ordination. Similarly, without residual term for the latent variables and with predictor variables, the model simplifies to a reduced rank regression (RRR) and is similar to the popular constrained ordination methods CCA and RDA.

In studies of species distribution modelling, the bioclimatic variables often used are collinear (Júnior & Nóbrega 2018), so that it may not always be possible to accurately estimate all parameters. In the first simulation study, performed competitively with MGLMs in estimating species responses, even while it included fewer parameters. Future research on the use and performance of RRR could attempt to further explore the dimensionality of species communities, which we here found to be lower than assumed by a MGLM. Though most ecologists consider between two and four dimensions for ordination (Halvorsen 2012), our results imply that at least five dimensions are necessary to accurately estimate species responses. However, the dimensionality of species responses might differ between communities based on e.g., the number of environmental gradients that underlay the structure of a community (see also Tobler *et al.* 2019).

In classical constrained ordination methods such as CCA and RDA, latent variables are assumed to be perfectly represented by predictor variables, so that those methods do not account for residual information unaccounted for by the predictors (ter Braak & Šmilauer 2015). In contrast, the approach presented here is capable of simultaneous unconstrained and constrained ordination though it requires having measured predictors, unlike in model-based unconstrained ordination (Hui *et al.* 2015). In the second simulation study in this article, we showed that if the predictors are unrelated to the true latent variables, so that the true model is that of an unconstrained ordination, CCA and RDA are unable to retrieve the latent variables and species responses, in contrast to the proposed GLLVM. In reality, it can often be unclear which predictors represent the latent variables, so that accounting for additional residual information can be important. Though CCA and RDA do not account for residual information in the ordination explicitly, our results suggest that the use of WA scores sufficiently mitigates that deficiency. WA scores can be considered minimally constrained, unlike LC scores (Palmer 1993), which have shown to not be sufficiently robust for inference (McCune 1997). Accounting for residual information from species responses addresses the concern shared by community ecologists over discarding ecological gradients that the ecologist is unaware of (Palmer 1993; Økland 1996; McCune 1997; ter Braak & Šmilauer 2015).

We demonstrated how to apply model-based ordination with constrained latent variables using two example datasets, one of Swiss alpine plants (D'Amen *et al.* 2017) and another of vascular plants in Norway. In those instances, the residual variation unaccounted for by the predictor variables demonstrated the need to account for residual variation in community ecological studies using dimension reduction techniques. We assessed importance of the predictors in the constrained ordination using a semi-partial $R^2_{\boldsymbol{B}}$ (Edwards *et al.*

2008), which can be calculated for our proposed model-based constrained ordination irrespective of whether the residual term is included or not (though omitting the residual term will affect the magnitude of the $R^2_{\boldsymbol{B}}$ statistic). In the first example, elevation and slope were the most important predictors in explaining the latent variables for the Swiss alpine plants dataset according to their semi-partial $R^2_{\boldsymbol{B}}$ values, and calcium for the Levanger grasslands dataset. For the Levanger grasslands dataset, the measured soil variables included as predictors did not provide a good representation of the latent variables, as indicated by a low $R^2_{\boldsymbol{B}}$ value for the entire model and by the parameter estimates and their corresponding Wald-tests for the predictor slopes $\boldsymbol{B}$.

Model-based ordination with constrained latent variables provides a suitable alternative for the modelling of multiple species responses with or without residual term or other random-effects (such as random row-intercepts). Our proposed approach provides access to standard tools for statistical inference such as statistical uncertainties for parameters estimates, model-selection tools, p-values related to a Wald-statistic that can be used to determine significance of the effect of predictors, and more such as residual diagnostics, all of which are available as part of the `gllvm` `R`-package (Niku *et al.* 2020), including a vignette that demonstrates the use of the proposed method. To conclude, the method presented here provides an extended version for various types of multivariate analyses, including fixed-effects constrained ordination, partial constrained ordination, MGLMs, and in general has merit for the ordering of sites and species.

Acknowledgements

The authors would like to thank Synnøve Nordal Grenne, Liv S. Nilsen and Line Rosef, who together with Knut Anders Hovstad collected the Levanger grasslands dataset, which is available on Figshare to download (https://doi.org/10.6084/m9.figshare.15143937.v1). The authors would like to thank Cajo ter Braak for providing the extended hunting spiders dataset, used in the first simulation study. Jenni Niku kindly helped improving the software implementation for the `gllvm` `R`-package. B.V. was supported by a scholarship from the Research Council of Norway (grant number 272408/F40). F.K.C.H was supported by an Australia Research Council Discovery Fellowship (grant number DE200100435).

Authors contributions

B.V. conceived the ideas. B.V. and F.K.C.H designed the methodology. All authors contributed to the writing, reviewing and editing of the draft and gave final approval for publication.

References

Anderson, T.W. (1951). Estimating Linear Restrictions on Regression Coefficients for Multivariate Normal Distributions. *The Annals of Mathematical Statistics*, **22**, 327–351. Retrieved June 29, 2021, from https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-22/issue-3/Estimating-Linear-Restrictions-on-Regression-Coefficients-for-Multivariate-Normal-Distributions/10.1214/aoms/1177729580.full

Booth, T.H. (2018). Why understanding the pioneering and continuing contributions of BIOCLIM to species distribution modelling is important. *Austral Ecology*, **43**, 852–860. Retrieved June 29, 2021, from https://onlinelibrary.wiley.com/doi/abs/10.1111/aec.12628

Booth, J.G. & Hobert, J.P. (1998). Standard Errors of Prediction in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, **93**, 262–272. Retrieved June 15, 2021, from https://www.tandfonline.com/doi/abs/10.1080/01621459.1998.10474107

Burnham, K.P. & Anderson, D.R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edn. Springer-Verlag, New York. Retrieved July 5, 2021, from https://www.springer.com/gp/book/9780387953649

D'Amen, M., Mod, H.K., Gotelli, N.J. & Guisan, A. (2017). Disentangling biotic interactions, environmental filters, and dispersal limitation as drivers of species co-occurrence. *Dryad*. Retrieved June 29, 2020, from http://datadryad.org/stash/dataset/doi:10.5061/dryad.8mv11

D'Amen, M., Mod, H.K., Gotelli, N.J. & Guisan, A. (2018). Disentangling biotic interactions, environmental filters, and dispersal limitation as drivers of species co-occurrence. *Ecography*, **41**, 1233–1244. Retrieved June 9, 2020, from https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.03148

Damgaard, C., Hansen, R.R. & Hui, F.K.C. (2020). Model-based ordination of pin-point cover data: Effect of management on dry heathland. *Ecological Informatics*, **60**, 101155. Retrieved July 5, 2021, from https://www.sciencedirect.com/science/article/pii/S1574954120301059

Dunn, P.K. & Smyth, G.K. (1996). Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics*, **5**, 236–244. Retrieved from http://www.jstor.org/stable/1390802

Edwards, L.J., Muller, K.E., Wolfinger, R.D., Qaqish, B.F. & Schabenberger, O. (2008). An R2 statistic for fixed effects in the linear mixed model. *Statistics in Medicine*, **27**, 6137–6157. Retrieved July 5, 2021, from https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3429

Files, B.T., Strelioff, M. & Bonnevie, R. (2019). *Bayesian reduced-rank regression with stan*. ARMY RESEARCH LAB ADELPHI MD Playa Vista United States.

Halvorsen, R. (2012). A gradient analytic perspective on distribution modelling. *Sommerfeltia*, **35**, 1–165.

Hartig, F. (2021). *DHARMa: Residual diagnostics for hierarchical (multi-level / mixed) regression models.* Retrieved from https://CRAN.R-project.org/package=DHARMa

Hawinkel, S., Kerckhof, F.-M., Bijnens, L. & Thas, O. (2019). A unified framework for unconstrained and constrained ordination of microbiome read count data. *PLOS ONE*, **14**, e0205474. Retrieved July 5, 2021, from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0205474

Hijmans, R.J. (2020). *Raster: Geographic data analysis and modeling.* Retrieved from https://CRAN.R-project.org/package=raster

Hoegh, A. & Roberts, D.W. (2020). Evaluating and presenting uncertainty in model-based unconstrained ordination. *Ecology and Evolution*, **10**, 59–69. Retrieved July 5, 2021, from https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.5752

Hui, F.K.C. (2016). Boral – Bayesian Ordination and Regression Analysis of Multivariate Abundance Data in r. *Methods in Ecology and Evolution*, **7**, 744–750. Retrieved August 21, 2020, from https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12514

Hui, F.K.C. (2017). Model-based simultaneous clustering and ordination of multivariate abundance data in ecology. *Computational Statistics & Data Analysis*, **105**, 1–10. Retrieved July 5, 2021, from https://www.sciencedirect.com/science/article/pii/S0167947316301724

Hui, F.K.C., Taskinen, S., Pledger, S., Foster, S.D. & Warton, D.I. (2015). Model-based approaches to unconstrained ordination. *Methods in Ecology and Evolution*, **6**, 399–411. Retrieved April 24, 2020, from https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12236

Hui, F.K.C., Warton, D.I., Ormerod, J.T., Haapaniemi, V. & Taskinen, S. (2017). Variational Approximations for Generalized Linear Latent Variable Models. *Journal of Computational and Graphical Statistics*, **26**, 35–43. Retrieved April 24, 2020, from https://doi.org/10.1080/10618600.2016.1164708

Jaeger, B.C., Edwards, L.J., Das, K. & Sen, P.K. (2017). An R2 statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics*, **44**, 1086–1105. Retrieved July 7, 2021, from https://doi.org/10.1080/02664763.2016.1193725

Jongman, R., ter Braak, C. & van Tongeren, O. (Eds.). (1995). *Data analysis in community and landscape ecology.* Cambridge university press, Cambridge.

Júnior, P.D.M. & Nóbrega, C.C. (2018). Evaluating collinearity effects on species distribution models: An approach based on virtual species simulation. *PLOS ONE*, **13**, e0202403. Retrieved June 29, 2021, from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0202403

Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H. & Bell, B. (2016). TMB: Automatic Differentiation and Laplace Approximation. *J. Stat. Soft.*, **70**.

McCune, B. (1997). Influence of Noisy Environmental Data on Canonical Correspondence Analysis. *Ecology*, **78**, 2617–2623. Retrieved April 22, 2021, from https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/0012-9658

Niku, J. (2020). On modeling multivariate abundance data with generalized linear latent variable models. *JYU dissertations*. Retrieved June 30, 2021, from https://jyx.jyu.fi/handle/123456789/67735

Niku, J., Brooks, W., Herliansyah, R., Hui, F.K.C., Taskinen, S., Warton, D.I. & van der Veen, B. (2020). *Gllvm: Generalized linear latent variable models*. Retrieved from https://CRAN.R-project.org/package=gllvm

Niku, J., Hui, F.K.C., Taskinen, S. & Warton, D.I. (2019). Gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in r. *Methods in Ecology and Evolution*, **10**, 2173–2182. Retrieved May 13, 2020, from https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13303

Niku, J., Warton, D.I., Hui, F.K.C. & Taskinen, S. (2017). Generalized Linear Latent Variable Models for Multivariate Count and Biomass Data in Ecology. *JABES*, **22**, 498–522. Retrieved April 24, 2020, from https://doi.org/10.1007/s13253-017-0304-7

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. & Wagner, H. (2020). *Vegan: Community ecology package*. Retrieved from https://CRAN.R-project.org/package=vegan

Palmer, M.W. (1993). Putting Things in Even Better Order: The Advantages of Canonical Correspondence Analysis. *Ecology*, **74**, 2215–2230. Retrieved from http://www.jstor.org/stable/1939575

Pebesma, E.J. & Bivand, R.S. (2005). Classes and methods for spatial data in R. *R News*, **5**, 9–13. Retrieved from https://CRAN.R-project.org/doc/Rnews/

Peres-Neto, P.R. & Jackson, D.A. (2001). How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. *Oecologia*, **129**, 169–178. Retrieved August 8, 2020, from https://doi.org/10.1007/s004420100720

Rao, C.R. (1964). The Use and Interpretation of Principal Component Analysis in Applied Research. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, **26**, 329–358. Retrieved from http://www.jstor.org/stable/25049339

ter Braak, C.J. (1986). Canonical Correspondence Analysis: A New Eigenvector Technique for Multivariate Direct Gradient Analysis. *Ecology*, **67**, 1167–1179. Retrieved May 25, 2020, from https://esajournals.onlinelibrary.wiley.com/doi/abs/10.2307/1938672

ter Braak, C.J.F. (1987). *Unimodal models to relate species to environment*. PhD thesis thesis, Ter Braak, S.l. Retrieved July 5, 2021, from https://library.wur.nl/WebQuery/wurpubs/2436

Ter Braak, C.J.F. (1987). The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio*, **69**, 69–77. Retrieved July 11, 2021, from https://doi.org/10.1007/BF00038688

ter Braak, C. (1988). Partial canonical correspondence analysis. *Classification and related methods of data analysis: proceedings of the first conference of the International Federation of Classification Societies (IFCS), Technical University of Aachen, FRG, 29 June-1 July 1987*, 551–558.

ter Braak, C.J.F. & Prentice, I.C. (1988). A Theory of Gradient Analysis. *Advances in Ecological Research* (eds M. Begon, A.H. Fitter, E.D. Ford & A. Macfadyen), pp. 271–317. Academic Press. Retrieved July 24, 2020, from http://www.sciencedirect.com/science/article/pii/S006525040860183X

ter Braak, C.J.F. & Šmilauer, P. (2015). Topics in constrained and unconstrained ordination. *Plant Ecol*, **216**, 683–696. Retrieved June 9, 2021, from https://doi.org/10.1007/s11258-014-0356-5

Tobler, M.W., Kéry, M., Hui, F.K.C., Guillera-Arroita, G., Knaus, P. & Sattler, T. (2019). Joint species distribution models with species correlations and imperfect detection. *Ecology*, **100**, e02754. Retrieved July 12, 2021, from https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/ecy.2754

van der Aart, P. & Smeek-Enserink, N. (1975). Correlations between distributions of hunting spiders (Lycosidae, Ctenidae) and environmental characteristics in a dune area. *Netherlands Journal of Zoology*, **25**, 1–45.

van der Veen, B., Hui, F.K.C., Hovstad, K.A., Solbu, E.B. & O'Hara, R.B. (2021). Model-based ordination for species with unequal niche widths. *Methods in Ecology and Evolution*, **n/a**. Retrieved April 20, 2021, from https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13595

Wang, Y., Naumann, U., Wright, S.T. & Warton, D.I. (2012). Mvabund– an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution*, **3**, 471–474. Retrieved August 12, 2020, from https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.2041-210X.2012.00190.x

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

Warton, D.I. & Hui, F.K.C. (2017). The central role of mean-variance relationships in the analysis of multivariate abundance data: A response to Roberts (2017). *Methods in Ecology and Evolution*, **8**, 1408–1414. Retrieved July 12, 2021, from https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12843

Yee, T.W. (2004). A New Technique for Maximum-Likelihood Canonical Gaussian Ordination. *Ecological Monographs*, **74**, 685–701. Retrieved June 9, 2020, from https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/03-0078

Yee, T.W. (2014). Reduced-rank vector generalized linear models with two linear predictors. *Computational Statistics & Data Analysis*, **71**, 889–902. Retrieved July 5, 2021, from https://www.sciencedirect.com/science/article/pii/S0167947313000273

Yee, T.W. & Hastie, T.J. (2003). Reduced-rank vector generalized linear models. *Statistical Modelling*, **3**, 15–41. Retrieved June 9, 2021, from https://doi.org/10.1191/1471082X03st045oa

Zeng, Y., Zhao, H. & Wang, T. (2021). Model-Based Microbiome Data Ordination: A Variational Approximation Approach. *Journal of Computational and Graphical Statistics*, **0**, 1–13. Retrieved July 5, 2021, from https://doi.org/10.1080/10618600.2021.1882467

Økland, R.H. (1996). Are ordination and constrained ordination alternative or complementary strategies in general ecological studies? *Journal of Vegetation Science*, **7**, 289–292. Retrieved from http://www.jstor.org/stable/3236330

Model-based ordination with constrained latent variables

Bert van der Veen[123] Francis K.C. Hui[4] Knut A. Hovstad[53]
Robert B. O'Hara[23]

[1]Department of Landscape and Biodiversity, Norwegian Institute of Bioeconomy research, Trondheim, Norway
[2]Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway
[3]Centre of Biodiversity Dynamics, Norwegian University of Science and Technology, Trondheim, Norway
[4]Research School of Finance, Actuarial Studies and Statistics, The Australian National University, Canberra, Australia
[5]The Norwegian Biodiversity Information Centre, Trondheim, Norway

Appendix S1: Variance of products of parameters

Here, we derive confidence intervals for the reduced rank approximated slopes. Let our model be:

$$\eta_{ij} = \beta_{0j} + \boldsymbol{X}_i^\top \boldsymbol{B}\boldsymbol{\gamma_j} - \boldsymbol{X}_i^\top \boldsymbol{B}\boldsymbol{D}_j\boldsymbol{B}^\top \boldsymbol{X}_i. \tag{1}$$

A supportnig proof can be found in Bohrnstedt and Goldberger 2012 eq. 13 (https://www.tandfonline.com/doi/abs/10.1080/01621459.1969.10501069).

Linear term

For $\boldsymbol{B}\boldsymbol{\gamma}_j$, where $\text{vech}(\boldsymbol{B}) = \{b_{11}, b_{21} \dots b_{kq}\}$ and $\boldsymbol{\gamma}_j = \{\gamma_{1j}, \dots \gamma_{qj}\}$, so that $\{\boldsymbol{B}, \boldsymbol{\gamma}_j\} \sim \mathcal{N}(\begin{pmatrix} \hat{b_{11}} \\ \vdots \\ \hat{b}_{kq} \\ \hat{\gamma}_{1j} \\ \vdots \\ \hat{\gamma}_{qj} \end{pmatrix}, \boldsymbol{\Sigma})$, where

$\Delta b_{kq} = \hat{b}_{kq} - \text{E}(b_{kq})$ and similar for $\Delta\gamma_{qj}$ (and so e.g. $\text{E}(\Delta\gamma_{j1}^3)$ is zero).

$$
\begin{aligned}
\text{var}(\boldsymbol{b}_k\boldsymbol{\gamma}_j) &= \text{var}(\sum_{q=1}^{d} b_{kq}\gamma_{qj}) \\
&= \sum_{q=1}^{d}\sum_{r=1}^{d} \text{cov}(b_{kq}\gamma_{qj}, b_{kr}\gamma_{rj}) \\
&\qquad \text{E}(b_{kq})\text{E}(b_{kr})\text{cov}(\gamma_{qj}, \gamma_{rj}) + \\
&\qquad \text{E}(b_{kq})\text{E}(\gamma_{rj})\text{cov}(\gamma_{qj}, b_{kr}) + \\
&= \sum_{q=1}^{d}\sum_{r=1}^{d} \text{E}(\gamma_{qj})\text{E}(b_{kr})\text{cov}(b_{kq}, \gamma_{rj}) + \\
&\qquad \text{E}(\gamma_{qj})\text{E}(\gamma_{rj})\text{cov}(b_{kq}, b_{kr}) + \\
&\qquad \text{E}(\Delta b_{kq}\Delta b_{kr}\Delta\gamma_{qj}\Delta\gamma_{rj}) - \text{cov}(b_{kq}, \gamma_{qj})\text{cov}(b_{kr}, \gamma_{rj}),
\end{aligned} \tag{2}
$$

By noting that

$$
\begin{aligned}
\text{E}(\Delta b_{kq}\Delta b_{kr}\Delta\gamma_{qj}\Delta\gamma_{rj}) = \text{cov}(b_{kq}, b_{kr})\text{cov}(\gamma_{qj}, \gamma_{rj}) + \\
\text{cov}(b_{kq}, \gamma_{qj})\text{cov}(b_{kr}, \gamma_{rj}) + \\
\text{cov}(b_{kq}, \gamma_{rj})\text{cov}(b_{kr}, \gamma_{qj}),
\end{aligned} \tag{3}
$$

we have the final solution

$$
\begin{aligned}
\text{var}(\boldsymbol{b}_k\boldsymbol{\gamma}_j) = \sum_{q=1}^{d}\sum_{r=1}^{d} \\
\hat{b}_{kq}\hat{b}_{kr}\text{cov}(\gamma_{qj}, \gamma_{rj}) + \\
\hat{\gamma}_{rj}\hat{b}_{kq}\text{cov}(\gamma_{qj}, b_{kr}) + \\
\hat{\gamma}_{qj}\hat{b}_{kr}\text{cov}(\gamma_{rj}, b_{kq}) + \\
\hat{\gamma}_{qj}\hat{\gamma}_{rj}\text{cov}(b_{kq}, b_{kr}) + \\
\text{cov}(b_{kq}, b_{kr})\text{cov}(\gamma_{qj}, \gamma_{rj}) + \\
\text{cov}(\gamma_{rj}, b_{kq})\text{cov}(\gamma_{qj}, b_{kr})
\end{aligned} \tag{4}
$$

Quadratic term

For the quadratic term, we recall that we are additionally interested in the variance of the estimator for $\boldsymbol{B}\boldsymbol{D}_j\boldsymbol{B}^\top$.

$$\begin{aligned}
\mathrm{var}(\boldsymbol{B}\boldsymbol{D}_j\boldsymbol{B}^\top) &= \mathrm{var}(\sum_{q=1}^{d} b_{kq}b_{ql}d_{qj}) \\
&= \sum_{q=1}^{d}\sum_{q=1}^{d} \mathrm{cov}(b_{kq}b_{ql}d_{qj}, b_{kr}b_{rl}d_{rj}) \\
&= \sum_{q=1}^{d}\sum_{q=1}^{d} \mathrm{E}(b_{kq}b_{ql}d_{qj}b_{kr}b_{rl}d_{rj}) - \mathrm{E}(b_{kq}b_{ql}d_{qj})\mathrm{E}(b_{kr}b_{rl}d_{rj}).
\end{aligned} \tag{5}$$

The first term in equation (5) is the 6th order moment of the multivariate normal distribution, and the second a product of two (non-central) third order moments. As such, first of the second terms has the solution:

$$\begin{aligned}
\mathrm{E}[b_{kq}b_{ql}d_{qj}] &= \mathrm{E}[\{\Delta b_{kq} + \mathrm{E}(b_{kq})\}\{\Delta b_{ql} + \mathrm{E}(b_{ql})\}\{\Delta d_{qj} + \mathrm{E}(d_{qj})\}] \\
&= \begin{array}{l}
\mathrm{E}[\Delta b_{kq}\Delta b_{ql}\Delta d_{qj} + \Delta b_{kq}\Delta b_{ql}\mathrm{E}\{d_{qj}\}+ \\
\Delta b_{kq}\mathrm{E}\{b_{ql}\}\Delta d_{qj} + \Delta b_{kq}\mathrm{E}\{b_{ql}\}\mathrm{E}\{d_{qj}\}+ \\
\mathrm{E}\{b_{kq}\}\Delta b_{ql}\Delta d_{qj} + \mathrm{E}\{b_{kq}\}\Delta b_{ql}\mathrm{E}\{d_{qj}\}+ \\
\mathrm{E}\{b_{kq}\}\mathrm{E}\{b_{ql}\}\Delta d_{qj} + \mathrm{E}\{b_{kq}\}\mathrm{E}\{b_{ql}\}\mathrm{E}\{d_{qj}\}]
\end{array} \\
&= \mathrm{E}[\Delta b_{kq}\Delta b_{ql}]\mathrm{E}[d_{jl}] + \mathrm{E}[\Delta b_{kq}\Delta d_{jl}]\mathrm{E}[b_{ql}] + \mathrm{E}[\Delta b_{ql}\Delta d_{jl}]\mathrm{E}[b_{kq}] \\
&= \mathrm{E}[d_{jl}]\mathrm{cov}(b_{kq}, b_{ql}) + \mathrm{E}[b_{ql}]\mathrm{cov}(b_{kq}, d_{jl}) + \mathrm{E}[b_{kq}]\mathrm{cov}(b_{ql}, d_{jl}).
\end{aligned} \tag{6}$$

The solution for the first term in equation (5) is calculated in a similar fashion:

$$\begin{aligned}
\mathrm{E}[b_{kq}b_{ql}b_{kr}b_{rl}d_{qj}d_{rj}] &= \mathrm{E}[\{\Delta b_{kq} + \mathrm{E}(b_{kq})\}\{\Delta b_{ql} + \mathrm{E}(b_{ql})\}\{\Delta b_{kr} + \mathrm{E}(b_{kr})\}\{\Delta b_{rl} + \mathrm{E}(b_{rl})\}\{\Delta d_{qj}\mathrm{E}(d_{qj})\}\{\Delta d_{rj} + \mathrm{E}(d_{rj})\}] \\
&= \mathrm{E}(\Delta b_{kq}\Delta b_{ql}\Delta b_{kr}\Delta b_{rl}\Delta d_{qj}\Delta d_{rj}) + \mathrm{E}(b_{ql})\mathrm{E}(b_{kq})\mathrm{E}(\Delta b_{kr}\Delta b_{rl}\Delta d_{qj}\Delta d_{rj})+ \\
&\mathrm{E}(b_{ql})\mathrm{E}(b_{kr})\mathrm{E}(\Delta b_{kq}\Delta b_{rl}\Delta d_{qj}\Delta d_{rj}) + \mathrm{E}(b_{kq})\mathrm{E}(b_{kr})\mathrm{E}(\Delta b_{ql}\Delta b_{rl}\Delta d_{qj}\Delta d_{rj})+ \\
&\mathrm{E}(b_{ql})\mathrm{E}(b_{rl})\mathrm{E}(\Delta b_{kq}\Delta b_{kr}\Delta d_{qj}\Delta d_{rj}) + \mathrm{E}(b_{kq})\mathrm{E}(b_{rl})\mathrm{E}(\Delta b_{ql}\Delta b_{kr}\Delta d_{qj}\Delta d_{rj})+ \\
&\mathrm{E}(b_{kr})\mathrm{E}(b_{rl})\mathrm{E}(\Delta b_{kq}\Delta b_{ql}\Delta d_{qj}\Delta d_{rj}) + \mathrm{E}(b_{ql})\mathrm{E}(d_{qj})\mathrm{E}(\Delta b_{kq}\Delta b_{kr}\Delta b_{rl}\Delta d_{rj})+ \\
&\mathrm{E}(b_{kq})\mathrm{E}(d_{qj})\mathrm{E}(\Delta b_{ql}\Delta b_{kr}\Delta b_{rl}\Delta d_{rj}) + \mathrm{E}(b_{kr})\mathrm{E}(d_{qj})\mathrm{E}(\Delta b_{kq}\Delta b_{ql}\Delta b_{rl}\Delta d_{rj})+ \\
&\mathrm{E}(b_{rl})\mathrm{E}(d_{qj})\mathrm{E}(\Delta b_{kq}\Delta b_{ql}\Delta b_{kr}\Delta d_{rj}) + \mathrm{E}(b_{ql})\mathrm{E}(d_{rj})\mathrm{E}(\Delta b_{kq}\Delta b_{kr}\Delta b_{rl}\Delta d_{qj})+ \\
&\mathrm{E}(b_{kq})\mathrm{E}(d_{rj})\mathrm{E}(\Delta b_{ql}\Delta b_{kr}\Delta b_{rl}\Delta d_{qj}) + \mathrm{E}(b_{kr})\mathrm{E}(d_{rj})\mathrm{E}(\Delta b_{kq}\Delta b_{ql}\Delta b_{rl}\Delta d_{qj})+
\end{aligned}$$

$$
\begin{aligned}
&\mathrm{E}(b_{rl})\mathrm{E}(d_{rj})\mathrm{E}(\Delta b_{kq}\Delta b_{ql}\Delta b_{kr}\Delta d_{qj}) + \mathrm{E}(d_{qj})\mathrm{E}(d_{rj})\mathrm{E}(\Delta b_{kq}\Delta b_{ql}\Delta b_{kr}\Delta b_{rl}) + \\
&\mathrm{E}(b_{kq})\mathrm{E}(b_{ql})\mathrm{E}(b_{kr})\mathrm{E}(b_{rl})\mathrm{E}(\Delta d_{qj}\Delta d_{rj}) + \mathrm{E}(b_{kq})\mathrm{E}(b_{ql})\mathrm{E}(b_{kr})\mathrm{E}(d_{qj})\mathrm{E}(\Delta b_{rl}\Delta d_{rj}) + \\
&\mathrm{E}(b_{kq})\mathrm{E}(b_{ql})\mathrm{E}(b_{rl})\mathrm{E}(d_{qj})\mathrm{E}(\Delta b_{kr}\Delta d_{rj}) + \mathrm{E}(b_{ql})\mathrm{E}(b_{kr})\mathrm{E}(b_{rl})\mathrm{E}(d_{qj})\mathrm{E}(\Delta b_{kq}\Delta d_{rj}) + \\
&\mathrm{E}(b_{kq})\mathrm{E}(b_{kr})\mathrm{E}(b_{rl})\mathrm{E}(d_{qj})\mathrm{E}(\Delta b_{ql}\Delta d_{rj}) + \mathrm{E}(b_{kq})\mathrm{E}(b_{ql})\mathrm{E}(b_{rl})\mathrm{E}(d_{rj})\mathrm{E}(\Delta b_{rl}\Delta d_{qj}) + \\
&\mathrm{E}(b_{kq})\mathrm{E}(b_{ql})\mathrm{E}(b_{rl})\mathrm{E}(d_{rj})\mathrm{E}(\Delta b_{kr}\Delta d_{qj}) + \mathrm{E}(b_{ql})\mathrm{E}(b_{kr})\mathrm{E}(b_{rl})\mathrm{E}(d_{rj})\mathrm{E}(\Delta b_{kq}\Delta d_{qj}) + \\
&\mathrm{E}(b_{kq})\mathrm{E}(b_{kr})\mathrm{E}(b_{rl})\mathrm{E}(d_{rj})\mathrm{E}(\Delta b_{ql}\Delta d_{qj}) + \mathrm{E}(b_{kq})\mathrm{E}(b_{ql})\mathrm{E}(d_{qj})\mathrm{E}(d_{rj})\mathrm{E}(\Delta b_{kr}\Delta b_{rl}) + \\
&\mathrm{E}(b_{ql})\mathrm{E}(b_{kr})\mathrm{E}(d_{qj})\mathrm{E}(d_{rj})\mathrm{E}(\Delta b_{kq}\Delta b_{rl}) + \mathrm{E}(b_{kq})\mathrm{E}(b_{kr})\mathrm{E}(d_{qj})\mathrm{E}(d_{rj})\mathrm{E}(\Delta b_{ql}\Delta b_{rl}) + \\
&\mathrm{E}(b_{ql})\mathrm{E}(b_{rl})\mathrm{E}(d_{qj})\mathrm{E}(d_{rj})\mathrm{E}(\Delta b_{kq}\Delta b_{kr}) + \mathrm{E}(b_{kq})\mathrm{E}(b_{rl})\mathrm{E}(d_{qj})\mathrm{E}(d_{rj})\mathrm{E}(\Delta b_{ql}\Delta b_{kr}) + \\
&\mathrm{E}(b_{kr})\mathrm{E}(b_{rl})\mathrm{E}(d_{qj})\mathrm{E}(d_{rj})\mathrm{E}(\Delta b_{kq}\Delta b_{ql}) + \mathrm{E}(b_{kq})\mathrm{E}(b_{ql})\mathrm{E}(b_{kr})\mathrm{E}(b_{rl})\mathrm{E}(d_{qj})\mathrm{E}(d_{rj}) \\
&= \mathrm{E}(\Delta b_{kq}\Delta b_{ql}\Delta b_{kr}\Delta b_{rl}\Delta d_{qj}\Delta d_{rj}) + \\
&\mathrm{E}(b_{kq})\mathrm{E}(b_{ql})\{\mathrm{cov}(b_{kr},b_{rl})\mathrm{cov}(d_{qj},d_{rj}) + \mathrm{cov}(b_{kr},d_{qj})\mathrm{cov}(b_{rl},d_{rj}) + \mathrm{cov}(b_{kr},d_{rj})\mathrm{cov}(b_{rl},d_{qj})\} + \\
&\mathrm{E}(b_{ql})\mathrm{E}(b_{kr})\{\mathrm{cov}(b_{kq},b_{rl})\mathrm{cov}(d_{qj},d_{rj}) + \mathrm{cov}(b_{kq},d_{qj})\mathrm{cov}(b_{rl},d_{rj}) + \mathrm{cov}(b_{kq},d_{rj})\mathrm{cov}(b_{rl},d_{qj})\} + \\
&\mathrm{E}(b_{kq})\mathrm{E}(b_{kr})\{\mathrm{cov}(b_{ql},b_{rl})\mathrm{cov}(d_{qj},d_{rj}) + \mathrm{cov}(b_{ql},d_{qj})\mathrm{cov}(b_{rl},d_{rj}) + \mathrm{cov}(b_{ql},d_{rj})\mathrm{cov}(b_{rl},d_{qj})\} + \\
&\mathrm{E}(b_{ql})\mathrm{E}(b_{rl})\{\mathrm{cov}(b_{kq},b_{kr})\mathrm{cov}(d_{qj},d_{rj}) + \mathrm{cov}(b_{kq},d_{qj})\mathrm{cov}(b_{kr},d_{rj}) + \mathrm{cov}(b_{kq},d_{rj})\mathrm{cov}(b_{kr},d_{qj})\} + \\
&\mathrm{E}(b_{kq})\mathrm{E}(b_{rl})\{\mathrm{cov}(b_{ql},b_{kr})\mathrm{cov}(d_{qj},d_{rj}) + \mathrm{cov}(b_{ql},d_{qj})\mathrm{cov}(b_{kr},d_{rj}) + \mathrm{cov}(b_{ql},d_{rj})\mathrm{cov}(b_{kr},d_{qj})\} + \\
&\mathrm{E}(b_{kr})\mathrm{E}(b_{rl})\{\mathrm{cov}(b_{kq},b_{ql})\mathrm{cov}(d_{qj},d_{rj}) + \mathrm{cov}(b_{kq},d_{qj})\mathrm{cov}(b_{ql},d_{rj}) + \mathrm{cov}(b_{kq},d_{rj})\mathrm{cov}(b_{ql},d_{qj})\} + \\
&\mathrm{E}(b_{ql})\mathrm{E}(d_{qj})\{\mathrm{cov}(b_{kq},b_{kr})\mathrm{cov}(b_{rl},d_{rj}) + \mathrm{cov}(b_{kq},b_{rl})\mathrm{cov}(b_{kr},d_{rj}) + \mathrm{cov}(b_{kq},d_{rj})\mathrm{cov}(b_{kr},b_{rl})\} + \\
&\mathrm{E}(b_{kq})\mathrm{E}(d_{qj})\{\mathrm{cov}(b_{ql},b_{kr})\mathrm{cov}(b_{rl},d_{rj}) + \mathrm{cov}(b_{ql},b_{rl})\mathrm{cov}(b_{kr},d_{rj}) + \mathrm{cov}(b_{ql},d_{rj})\mathrm{cov}(b_{kr},b_{rl})\} + \\
&\mathrm{E}(b_{kr})\mathrm{E}(d_{qj})\{\mathrm{cov}(b_{kq},b_{ql})\mathrm{cov}(b_{rl},d_{rj}) + \mathrm{cov}(b_{kq},b_{rl})\mathrm{cov}(b_{ql},d_{rj}) + \mathrm{cov}(b_{kq},d_{rj})\mathrm{cov}(b_{ql},b_{rl})\} + \\
&\mathrm{E}(b_{rl})\mathrm{E}(d_{qj})\{\mathrm{cov}(b_{kq},b_{ql})\mathrm{cov}(b_{kr},d_{rj}) + \mathrm{cov}(b_{kq},b_{kr})\mathrm{cov}(b_{ql},d_{rj}) + \mathrm{cov}(b_{kq},d_{rj})\mathrm{cov}(b_{ql},b_{kr})\} + \\
&\mathrm{E}(b_{ql})\mathrm{E}(d_{rj})\{\mathrm{cov}(b_{kq},b_{kr})\mathrm{cov}(b_{rl},d_{qj}) + \mathrm{cov}(b_{kq},b_{rl})\mathrm{cov}(b_{kr},d_{qj}) + \mathrm{cov}(b_{kq},d_{qj})\mathrm{cov}(b_{kr},b_{rl})\} + \\
&\mathrm{E}(b_{kq})\mathrm{E}(d_{rj})\{\mathrm{cov}(b_{ql},b_{kr})\mathrm{cov}(b_{rl},d_{qj}) + \mathrm{cov}(b_{ql},b_{rl})\mathrm{cov}(b_{kr},d_{qj}) + \mathrm{cov}(b_{ql},d_{qj})\mathrm{cov}(b_{kr},b_{rl})\} + \\
&\mathrm{E}(b_{kr})\mathrm{E}(d_{rj})\{\mathrm{cov}(b_{kq},b_{ql})\mathrm{cov}(b_{rl},d_{qj}) + \mathrm{cov}(b_{kq},b_{rl})\mathrm{cov}(b_{ql},d_{qj}) + \mathrm{cov}(b_{kq},d_{qj})\mathrm{cov}(b_{ql},b_{rl})\} + \\
&\mathrm{E}(b_{rl})\mathrm{E}(d_{rj})\{\mathrm{cov}(b_{kq},b_{ql})\mathrm{cov}(b_{kr},d_{qj}) + \mathrm{cov}(b_{kq},b_{kr})\mathrm{cov}(b_{ql},d_{qj}) + \mathrm{cov}(b_{kq},d_{qj})\mathrm{cov}(b_{ql},b_{kr})\} + \\
&\mathrm{E}(d_{qj})\mathrm{E}(d_{rj})\{\mathrm{cov}(b_{kq},b_{ql})\mathrm{cov}(b_{kr},b_{rl}) + \mathrm{cov}(b_{kq},b_{kr})\mathrm{cov}(b_{ql},b_{rl}) + \mathrm{cov}(b_{kq},b_{rl})\mathrm{cov}(b_{ql},b_{kr})\} + \\
&\mathrm{E}(b_{kq})\mathrm{E}(b_{ql})\mathrm{E}(b_{kr})\mathrm{E}(b_{rl})\mathrm{cov}(d_{qj},d_{rj}) + \mathrm{E}(b_{kq})\mathrm{E}(b_{ql})\mathrm{E}(b_{kr})\mathrm{E}(d_{qj})\mathrm{cov}(b_{rl},d_{rj}) + \\
&\mathrm{E}(b_{kq})\mathrm{E}(b_{ql})\mathrm{E}(b_{rl})\mathrm{E}(d_{qj})\mathrm{cov}(b_{kr},d_{rj}) + \mathrm{E}(b_{ql})\mathrm{E}(b_{kr})\mathrm{E}(b_{rl})\mathrm{E}(d_{qj})\mathrm{cov}(b_{kq},d_{rj}) + \\
&\mathrm{E}(b_{kq})\mathrm{E}(b_{kr})\mathrm{E}(b_{rl})\mathrm{E}(d_{qj})\mathrm{cov}(b_{ql},d_{rj}) + \mathrm{E}(b_{kq})\mathrm{E}(b_{ql})\mathrm{E}(b_{rl})\mathrm{E}(d_{rj})\mathrm{cov}(b_{rl},d_{qj}) +
\end{aligned}
$$

$$\mathrm{E}(b_{kq})\mathrm{E}(b_{ql})\mathrm{E}(b_{rl})\mathrm{E}(d_{rj})\mathrm{cov}(b_{kr},d_{qj}) + \mathrm{E}(b_{ql})\mathrm{E}(b_{kr})\mathrm{E}(b_{rl})\mathrm{E}(d_{rj})\mathrm{cov}(b_{kq},d_{qj}) +$$
$$\mathrm{E}(b_{kq})\mathrm{E}(b_{kr})\mathrm{E}(b_{rl})\mathrm{E}(d_{rj})\mathrm{cov}(b_{ql},d_{qj}) + \mathrm{E}(b_{kq})\mathrm{E}(b_{ql})\mathrm{E}(d_{qj})\mathrm{E}(d_{rj})\mathrm{cov}(b_{kr},b_{rl}) +$$
$$\mathrm{E}(b_{ql})\mathrm{E}(b_{kr})\mathrm{E}(d_{qj})\mathrm{E}(d_{rj})\mathrm{cov}(b_{kq},b_{rl}) + \mathrm{E}(b_{kq})\mathrm{E}(b_{kr})\mathrm{E}(d_{qj})\mathrm{E}(d_{rj})\mathrm{cov}(b_{ql},b_{rl}) +$$
$$\mathrm{E}(b_{ql})\mathrm{E}(b_{rl})\mathrm{E}(d_{qj})\mathrm{E}(d_{rj})\mathrm{cov}(b_{kq},b_{kr}) + \mathrm{E}(b_{kq})\mathrm{E}(b_{rl})\mathrm{E}(d_{qj})\mathrm{E}(d_{rj})\mathrm{cov}(b_{ql},b_{kr}) +$$
$$\mathrm{E}(b_{kr})\mathrm{E}(b_{rl})\mathrm{E}(d_{qj})\mathrm{E}(d_{rj})\mathrm{cov}(b_{kq},b_{ql}) + \mathrm{E}(b_{kq})\mathrm{E}(b_{ql})\mathrm{E}(b_{kr})\mathrm{E}(b_{rl})\mathrm{E}(d_{qj})\mathrm{E}(d_{rj})$$

Appendix S2: Quadratic response model

As in van der Veen *et al.* (2021), the constrained GLLVM can also be extended with quadratic terms. Recall from equation (1) in the main text that a GLLVM with constrained latent variables is formulated as:

$$g\{\mathrm{E}(y_{ij}|\boldsymbol{x}_{lv,i},\boldsymbol{\epsilon}_i)\} = \beta_{0j} + \boldsymbol{x}_{lv,i}^\top \boldsymbol{B}\boldsymbol{\gamma}_j + \boldsymbol{\epsilon}_i^\top \boldsymbol{\gamma}_j. \tag{7}$$

Now, with quadratic response model this model becomes:

$$g\{\mathrm{E}(y_{ij}|\boldsymbol{x}_{lv,i},\boldsymbol{\epsilon}_i)\} = \beta_{0j} + \boldsymbol{x}_{lv,i}^\top \boldsymbol{B}\boldsymbol{\gamma}_j - \boldsymbol{x}_{lv,i}^\top \boldsymbol{B}\boldsymbol{D}_j\boldsymbol{B}^\top \boldsymbol{x}_{lv,i} + \boldsymbol{\epsilon}_i^\top \boldsymbol{\gamma}_j - \boldsymbol{\epsilon}_i^\top \boldsymbol{D}_j\boldsymbol{\epsilon}_i - 2\boldsymbol{\epsilon}_i^\top \boldsymbol{D}_j\boldsymbol{B}^\top \boldsymbol{x}_{lv,i}, \tag{8}$$

where $\boldsymbol{D}_j$ is a positive-definite diagonal matrix that contains the quadratic coefficients for each species. Clearly this more complex model is heavily parametrized, and care should be taken to ensure that sufficient information is present in a dataset to accurately estimate all parameters. In fact, Yee (2004) encountered numerical instability when fitting the fixed-effects equivalent of this model i.e., when $\boldsymbol{\epsilon}_i = \boldsymbol{0}$, so that numerical issues are expected when attempting to fit this extended version too. Species optima, tolerances, and gradient length, for the latent variable $\boldsymbol{z}_i$ can be calculated similarly as in van der Veen *et al.* (2021). For the fixed-effects terms alone, species optima are calculated as $(2\boldsymbol{B}\boldsymbol{D}_j\boldsymbol{B}^\top)^{-1}(\boldsymbol{B}\boldsymbol{\gamma}_j - 2\boldsymbol{B}\boldsymbol{D}_j\boldsymbol{\epsilon}_i)$, where the second term is zero for constrained ordination without residual term (i.e. reduced rank regression). Similarly, for constrained ordination with and without residual term, species tolerances are retrieved from the diagonal of the matrix $(2\boldsymbol{B}\boldsymbol{D}_j\boldsymbol{B}^\top)^{-1}$, which is here the (reduced rank) covariance matrix of the fixed-effects ecological niche. Clearly, unlike in the quadratic response model for unconstrained latent variables as in van der Veen *et al.* (2021), this matrix is not diagonal, so that off-diagonal entries represent the (dis)similarity of a species response to two predictors. Though the precision matrix is singular as a consequence of modelling it in a reduced-rank form, and tolerances are as such more difficult to retrieve, a generalized inverse calculation can be employed to retrieve the covariance matrix. Then, per species a correlation matrix can be calculated, where a one means that a species responds the same to two environmental variables, and a zero that a species

response to two predictors is independent (i.e. the response to one predictor is not affected by the response to another).

Yee (2004) refers to this method as "Canonical Gaussian Ordination", but later re-named it to "Quadratic Constrained Ordination" instead, due to potential confusion with the Gaussian probability density function (Yee 2015). When assuming $\boldsymbol{D}_j$ to be the same for all species and latent variables, this is indeed the more exact version of the method presented by ter Braak (1986), although for a larger variety of data types as the response distribution can be adapted flexibly.

Due to the hierarchical formulation for the latent variable $\boldsymbol{z}_i$ the model becomes quite complex. In terms of univariate models for each species, this model includes a second order polynomial, with interactions between the linear forms of predictors. Most ecologists do not tend to fit interaction terms for polynomials in univariate statistics, or even quadratic curves without interactions for that matter (Austin 2007), so that most commonly the quadratic term here is assumed to be diagonal in ecological studies. In general, constrained quadratic ordination should be considered relative to the full rank model it approximates with d dimensions. Clearly, this is a complex model due to the non-independence of a species response to various environmental predictors. Not surprisingly, Yee (2004) reports numerical problems with this rather complex model, which we also experienced with the implementation of constrained quadratic ordination in the `gllvm` `R`-package.

Potentially, the numerical issues are caused by the high degree of complexity in the model, so that it can quickly overfit for small datasets. A more realistic assumption for many ecological datasets in $\boldsymbol{D}_j = \boldsymbol{D}$, i.e. the same quadratic coefficient for species on a latent variable (Yee 2004; van der Veen *et al.* 2021). Alternatively, quadratic species responses can be accommodated with a simpler structure in the linear response model, up to the sign constraint for the quadratic coefficients by assuming $\boldsymbol{z}_i = \boldsymbol{B}_1^\top \boldsymbol{x}_{lv,i} + (\text{diag}(\boldsymbol{x}_{lv,i})\boldsymbol{B}_2)^\top \boldsymbol{x}_{lv,i}$ where $\boldsymbol{B} = \{\boldsymbol{B}_1, \boldsymbol{B}_2\}$ and if we assume $\boldsymbol{\gamma}_j = \{\boldsymbol{\theta}_j, \text{diag}(\boldsymbol{D})_j\}$:

$$\begin{aligned} g\{\text{E}(y_{ij}|\boldsymbol{x}_{lv,i})\} &= \beta_{0j} + \boldsymbol{z}_i^\top \boldsymbol{\gamma}_j \\ \boldsymbol{z}_i^\top &= \boldsymbol{B}_1^\top \boldsymbol{x}_{lv,i} + (\text{diag}(\boldsymbol{x}_{lv,i})\boldsymbol{B}_2)^\top \boldsymbol{x}_{lv,i} \\ &= \beta_{0j} + \boldsymbol{x}_{lv}^\top \boldsymbol{B}_1 \boldsymbol{\theta}_j + \text{diag}(\boldsymbol{x}_{lv,i})\boldsymbol{B}_2 \boldsymbol{D}_j \boldsymbol{x}_{lv,i}, \end{aligned} \tag{9}$$

corresponding to a standard (multivariate) quadratic regression without interaction terms and sign constraints.

Appendix S3: Random slopes formulation

For the same model as before,

$$g\{\mathrm{E}(y_{ij}|\boldsymbol{x}_{lv,i},\boldsymbol{\epsilon}_i)\} = \beta_{0j} + \boldsymbol{x}_{lv,i}^\top \boldsymbol{B}\boldsymbol{\gamma}_j + \boldsymbol{\epsilon}_i^\top \boldsymbol{\gamma}_j, \tag{10}$$

we can instead assume that the slopes for the predictors $\boldsymbol{B}$ are random effects, so that for predictor $K = 1 \dots K$ and latent variable $q = 1 \dots d$, we have $b_{kq} \sim \mathcal{N}(0,1)$. This type of random slope effect can serve to induce shrinkage on the constrained ordination, similar as in ridge regression.

With $\boldsymbol{\epsilon}_i \sim \mathcal{N}\{\boldsymbol{0}, \mathrm{diag}(\boldsymbol{\sigma})\}$, and by noting that $b_{kq}x_{ik} \sim \mathcal{N}(0, x_{ik}^2)$, we can now continue to determine the distribution of

$$\boldsymbol{z}_i = (\boldsymbol{B}^\top \boldsymbol{X}_i + \boldsymbol{\epsilon}_i) \sim \mathcal{N}\Bigg\{\boldsymbol{0}, \begin{pmatrix} \sigma_1^2 + \sum_{k=1}^{K} x_{ik}^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_d^2 + \sum_{k=1}^{K} x_{ik}^2 \end{pmatrix}\Bigg\}, \tag{11}$$

so that it becomes straightforward to apply the Laplace approximation or the Variational approximation, similar as in Niku *et al.* (2017) and Hui *et al.* (2017) or van der Veen *et al.* (2021). For example, with a variational distribution for the random slopes $q(\boldsymbol{b}_k) = \mathcal{N}(\boldsymbol{v}_k, \boldsymbol{V}_k)$, where $\boldsymbol{v}_k$ is a d-sized vector of variational means per predictor and $\boldsymbol{V}_k$ is a $d \times d$ covariance matrix, with $q(\boldsymbol{\epsilon}_i) = \mathcal{N}(\boldsymbol{a}_i, \boldsymbol{A}_i)$, and with $q(\boldsymbol{B}, \boldsymbol{\epsilon}_i) = q(\boldsymbol{\epsilon}_i) \prod_{k=1}^{K} q(\boldsymbol{b}_k)$, i.e. independence of the random slopes and latent variables even for the variational distributions, we see that the variational distribution for the site scores becomes:

$$\boldsymbol{z}_i \sim \mathcal{N}\Bigg(\boldsymbol{a}_i + \sum_{k=1}^{K} \boldsymbol{v}_k x_{ik},\ \boldsymbol{A}_i + \sum_{k=1}^{K} x_{ik}^2 \boldsymbol{V}_k\Bigg), \tag{12}$$

so that the same calculations can be applied for integration of a constrained model with random slopes, as for an unconstrained ordination.

Appendix S4: Conditional Mean Squared Error of Predictions

An approximate conditional mean squared errors of predictions (CMSEPs) can be calculated for the site scores following Niku (2020) and Booth & Hobert (1998). For the constrained GLLVM, with $i = 1 \dots n$ sites and $j = 1 \dots p$ species as in (7) we denote the true site scores as $\boldsymbol{z}_i = \boldsymbol{x}_{lv,i}\boldsymbol{B} + \boldsymbol{\epsilon}_i$. Here we demonstrate the calculation using Variational Approximations (VA), though it is similar when using the Laplace approximation (Niku 2020). With VA, so that we have the variational distribution $\mathrm{q}(\boldsymbol{\epsilon}_i) \sim \mathcal{N}(\boldsymbol{a}_i, \boldsymbol{A}_i)$, the predicted site scores given the data, are $\boldsymbol{z}_{va,i} = \boldsymbol{x}_{lv,i}\boldsymbol{B} + \boldsymbol{a}_i$, and let $\hat{\boldsymbol{z}}_{va,i} = \boldsymbol{x}_{lv,i}\hat{\boldsymbol{B}} + \hat{\boldsymbol{a}}_i$ denote the estimated site

scores. Then, the CMSEP are given as

$$\begin{aligned}\text{CMSEP} &= \text{E}\{(\hat{\boldsymbol{z}}_{va,i} - \boldsymbol{z}_{va,i})(\hat{\boldsymbol{z}}_{va,i} - \boldsymbol{z}_{va,i})^\top | \boldsymbol{y}_i\} \\ &= \text{E}\{(\boldsymbol{z}_i - \boldsymbol{z}_{va,i})(\boldsymbol{z}_i - \boldsymbol{z}_{va,i})^\top | \boldsymbol{y}_i\} + \text{E}\{(\hat{\boldsymbol{z}}_{va,i} - \boldsymbol{z}_i)(\hat{\boldsymbol{z}}_{va,i} - \boldsymbol{z}_i)^\top | \boldsymbol{y}_i\}.\end{aligned}$$

Now, if we assume the true site score given the data to be equal to the true VA estimate, i.e. $\text{E}(\boldsymbol{z}_i|\boldsymbol{y}) = \boldsymbol{z}_{va,i}$, then the first term in the second line of equation (13) is $\text{cov}(\boldsymbol{z}_i|\boldsymbol{y}_i) \approx \boldsymbol{A}_i$. With a set of nuisance parameters $\text{vec}(\boldsymbol{B}, \dots \Theta)$, for the second term in equation (13), and specifically the part $\hat{\boldsymbol{z}}_{va,i} - \boldsymbol{z}_i = \boldsymbol{x}_{lv,i}(\hat{\boldsymbol{B}} - \boldsymbol{B}) + \hat{\boldsymbol{a}}_i - \boldsymbol{a}_i$. Using a Taylor series expansion: $\hat{\boldsymbol{a}}_i - \boldsymbol{a}_i \approx \left(\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{a}_i \partial \boldsymbol{a}_i^\top}\right)^{-1} \left(\frac{\partial \mathcal{L}}{\partial \boldsymbol{a}_i \partial \boldsymbol{\phi}^\top}\right)(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi})$.

Next, $\boldsymbol{x}_{lv,i}(\hat{\boldsymbol{B}} - \boldsymbol{B}) = Q(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi})$, for a $N_\phi \times Kd$ matrix $\boldsymbol{Q}$ with d replicates of $\boldsymbol{x}_{lv,i}$ on the diagonal, so that we have $\boldsymbol{x}_{lv,i}(\hat{\boldsymbol{B}} - \boldsymbol{B}) + \hat{\boldsymbol{a}}_i - \boldsymbol{a}_i \approx (Q + \left(\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{a}_i \partial \boldsymbol{a}_i^\top}\right)^{-1} \left(\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{a}_i \partial \boldsymbol{\phi}^\top}\right))(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}) \triangleq \boldsymbol{R}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi})$. Concluding,

$$\begin{aligned}\text{E}\{(\hat{\boldsymbol{z}}_{va,i} - \boldsymbol{z}_i)(\hat{\boldsymbol{z}}_{va,i} - \boldsymbol{z}_i)^\top | \boldsymbol{y}\} &\approx \text{E}\{\boldsymbol{R}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi})\boldsymbol{\mathcal{I}}_\phi^{-1}\boldsymbol{R}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi})^\top | \boldsymbol{y}\} \\ &\approx \boldsymbol{R}\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{R}^\top.\end{aligned}$$

Appendix S5: Residual covariance

For a model-based ordination with constrained latent variables as in equation (8), i.e. with the latent variable $\boldsymbol{z}_i = \boldsymbol{B}^\top \boldsymbol{x}_{lv,i} + \boldsymbol{\epsilon}_i$ and $\mathcal{N}(\boldsymbol{0}, \Sigma_{lv}$, where $\Sigma_{lv} = \boldsymbol{\sigma}^2 \text{I}_d)$, the residual covariance can be calculated straightforwardly, by noting that the residual term of the model in equation (8) is:

$$u_{ij} = \boldsymbol{\epsilon}_i^\top(\boldsymbol{\gamma}_j - 2\boldsymbol{x}_{lv,i}^\top \boldsymbol{B}\boldsymbol{D}_j) - \boldsymbol{\epsilon}_i^\top \boldsymbol{D}_j \boldsymbol{\epsilon}_i. \tag{13}$$

This form allows for a similar calculation as in van der Veen *et al.* (2021), though it should be noted that it depends on the observed predictor variables $\boldsymbol{x}_{lv,i}$. Thus, the elements of the residual covariance matrix for species j, k and sites i, l are :

$$
\begin{aligned}
\Sigma_{jl,ik} &= \operatorname{cov}\{u_{ij}, u_{kl}\}, \qquad i,k = 1\ldots n,\ j,l = 1\ldots p \\
&= \operatorname{cov}\{\boldsymbol{\epsilon}_i^\top(\boldsymbol{\gamma}_j - 2\boldsymbol{x}_{lv,i}^\top \boldsymbol{B}\boldsymbol{D}_j) - \boldsymbol{\epsilon}_i^\top \boldsymbol{D}_j\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_k^\top(\boldsymbol{\gamma}_l - 2\boldsymbol{x}_k^\top \boldsymbol{B}\boldsymbol{D}_l) - \boldsymbol{\epsilon}_k^\top \boldsymbol{D}_l\boldsymbol{\epsilon}_k\} \\
&= \operatorname{cov}\{\boldsymbol{\epsilon}_i^\top(\boldsymbol{\gamma}_j - 2\boldsymbol{x}_{lv,i}^\top \boldsymbol{B}\boldsymbol{D}_j), \boldsymbol{\epsilon}_k^\top(\boldsymbol{\gamma}_l - 2\boldsymbol{x}_k^\top \boldsymbol{B}\boldsymbol{D}_l)\} \\
&+ \operatorname{cov}\{\boldsymbol{\epsilon}_i^\top(\boldsymbol{\gamma}_j - 2\boldsymbol{x}_{lv,i}^\top \boldsymbol{B}\boldsymbol{D}_j), -\boldsymbol{\epsilon}_k^\top \boldsymbol{D}_l\boldsymbol{\epsilon}_k\} \\
&+ \operatorname{cov}\{-\boldsymbol{\epsilon}_i^\top \boldsymbol{D}_j\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_k^\top(\boldsymbol{\gamma}_l - 2\boldsymbol{x}_k^\top \boldsymbol{B}\boldsymbol{D}_l)\} \\
&+ \operatorname{cov}\{-\boldsymbol{\epsilon}_i^\top \boldsymbol{D}_j\boldsymbol{\epsilon}_i, -\boldsymbol{\epsilon}_k^\top \boldsymbol{D}_l\boldsymbol{\epsilon}_k\} \\
&= \boldsymbol{\gamma}_j^\top \Sigma_{lv}\boldsymbol{\gamma}_l + 2\operatorname{tr}(\boldsymbol{D}_j\Sigma_{lv}\boldsymbol{D}_k\Sigma_{lv}) - 2\boldsymbol{x}_i^\top \boldsymbol{B}\boldsymbol{D}_j\Sigma_{lv}\boldsymbol{\gamma}_l - 2\boldsymbol{x}_k^\top \boldsymbol{B}\boldsymbol{D}_l\Sigma_{lv}\boldsymbol{\gamma}_j + 4\boldsymbol{x}_{lv,i}^\top \boldsymbol{B}\boldsymbol{D}_j\Sigma_{lv}\boldsymbol{D}_l\boldsymbol{B}^\top\boldsymbol{x}_k,
\end{aligned}
\tag{14}
$$

where only the first two terms are applicable to the GLLVM with unconstrained latent variables and quadratic responses, and only the first term for a GLLVM with unconstrained latent variables and linear responses (i.e., where $\boldsymbol{D}_j = 0$).

Appendix S6: R-code for simulations

Code for the first simulation study

```
library(raster)
library(rworldmap)


# get world map
world <- getMap()
eu <- c("Austria", "Belgium", "Bulgaria", "Croatia", "Cyprus", "Czech Rep.",
    "Denmark", "Estonia", "Finland", "France", "Germany", "Greece", "Hungary",
    "Ireland", "Italy", "Latvia", "Lithuania", "Luxembourg", "Malta", "Netherlands",
    "Poland", "Portugal", "Romania", "Slovakia", "Slovenia", "Spain", "Sweden",
    "Switzerland", "Norway")



# Select only the index of states member of the E.U.
eu <- world[which(world$NAME %in% eu), ]
set.seed(1)
```

```r
coords <- spsample(eu, n = 1000, type = "random")
bioclim <- getData("worldclim", var = "bio", res = 10)
bioclim_eu <- crop(bioclim, extent(eu))
bioclim_extract <- extract(bioclim, coords)
bioclim_extract <- bioclim_extract[!apply(bioclim_extract, 1, function(x) any(is.na(x))),
    ]

n <- 100  #number of sites
p <- 12  #number of species

set.seed(1)
X <- mvtnorm::rmvnorm(n, rep(0, ncol(bioclim_extract)), cov(bioclim_extract))[,
    -7]  #simulate but remove collinear predictor
X <- scale(X)
colnames(X) <- colnames(bioclim_extract)[-7]

# read spider data, not provided with article!
data <- read.csv("Spiders_100cases_1985.csv")
data <- data[-1, -c(1:2)]
data <- apply(data, 2, as.integer)
mod <- manyglm(as.matrix(data) ~ X, family = "poisson")  #run model for spider data to get true values
set.seed(1)
mod$coefficients[1, ] <- runif(p, -1, 1)  #replace intercept
totres <- NULL
R <- 1000  #number of simulations

for (r in 1:R) {
    # generate data
    set.seed(r)
    y1 <- matrix(rpois(p * n, exp(eta)), ncol = p, nrow = n)
    set.seed(r)
    y2 <- matrix(rbinom(p * n, size = 1, prob = pnorm(eta)), ncol = p,
```

```
    nrow = n)
set.seed(r)
y3 <- matrix(rnorm(p * n, eta), ncol = p, nrow = n)

# fit multivariate GLMs
mod1 <- manyglm(y1 ~ X, family = "poisson")
mod2 <- manyglm(y2 ~ X, family = "binomial")
mod3 <- manylm(y3 ~ X)

# fit RRRs, making sure we get a model that converges. We start with
# most basic stating values

mod4 <- try(gllvm(y1, X = X, num.RR = 2, family = "poisson", sd.errors = F,
    maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
while (is.infinite(logLik(mod4))) {
    mod4 <- try(gllvm(y1, X = X, num.RR = 2, family = "poisson", sd.errors = F,
        maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
}
mod5 <- try(gllvm(y2, X = X, num.RR = 2, family = "binomial", sd.errors = F,
    maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
while (is.infinite(logLik(mod5))) {
    mod5 <- try(gllvm(y2, X = X, num.RR = 2, family = "binomial", sd.errors = F,
        maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
}
mod6 <- try(gllvm(y3, X = X, num.RR = 2, family = "gaussian", sd.errors = F,
    maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
while (is.infinite(logLik(mod6))) {
    mod6 <- try(gllvm(y3, X = X, num.RR = 2, family = "gaussian", sd.errors = F,
        maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
}
mod4a <- try(gllvm(y1, X = X, num.RR = 3, family = "poisson", sd.errors = F,
    maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
```

```
while (is.infinite(logLik(mod4a))) {
    mod4a <- try(gllvm(y1, X = X, num.RR = 3, family = "poisson", sd.errors = F,
        maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
}
mod5a <- try(gllvm(y2, X = X, num.RR = 3, family = "binomial", sd.errors = F,
    maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
while (is.infinite(logLik(mod5a))) {
    mod5a <- try(gllvm(y2, X = X, num.RR = 3, family = "binomial",
        sd.errors = F, maxit = 1e+05, max.iter = 1e+05, starting.val = "res"),
        silent = T)
}
mod6a <- try(gllvm(y3, X = X, num.RR = 3, family = "gaussian", sd.errors = F,
    maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
while (is.infinite(logLik(mod6a))) {
    mod6a <- try(gllvm(y3, X = X, num.RR = 3, family = "gaussian",
        sd.errors = F, maxit = 1e+05, max.iter = 1e+05, starting.val = "res"),
        silent = T)
}
mod4b <- try(gllvm(y1, X = X, num.RR = 4, family = "poisson", sd.errors = F,
    maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
while (is.infinite(logLik(mod4b))) {
    mod4b <- try(gllvm(y1, X = X, num.RR = 4, family = "poisson", sd.errors = F,
        maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
}
mod5b <- try(gllvm(y2, X = X, num.RR = 4, family = "binomial", sd.errors = F,
    maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
while (is.infinite(logLik(mod5b))) {
    mod5b <- try(gllvm(y2, X = X, num.RR = 4, family = "binomial",
        sd.errors = F, maxit = 1e+05, max.iter = 1e+05, starting.val = "res"),
        silent = T)
}
mod6b <- try(gllvm(y3, X = X, num.RR = 4, family = "gaussian", sd.errors = F,
```

```
    maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
while (is.infinite(logLik(mod6b))) {
    mod6b <- try(gllvm(y3, X = X, num.RR = 4, family = "gaussian",
        sd.errors = F, maxit = 1e+05, max.iter = 1e+05, starting.val = "res"),
        silent = T)
}
mod4c <- try(gllvm(y1, X = X, num.RR = 5, family = "poisson", sd.errors = F,
    maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
while (is.infinite(logLik(mod4c))) {
    mod4c <- try(gllvm(y1, X = X, num.RR = 5, family = "poisson", sd.errors = F,
        maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
}
mod5c <- try(gllvm(y2, X = X, num.RR = 5, family = "binomial", sd.errors = F,
    maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
while (is.infinite(logLik(mod5c))) {
    mod5c <- try(gllvm(y2, X = X, num.RR = 5, family = "binomial",
        sd.errors = F, maxit = 1e+05, max.iter = 1e+05, starting.val = "res"),
        silent = T)
}
mod6c <- try(gllvm(y3, X = X, num.RR = 5, family = "gaussian", sd.errors = F,
    maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
while (is.infinite(logLik(mod6c))) {
    mod6c <- try(gllvm(y3, X = X, num.RR = 5, family = "gaussian",
        sd.errors = F, maxit = 1e+05, max.iter = 1e+05, starting.val = "res"),
        silent = T)
}

mod4d <- try(gllvm(y1, X = X, num.RR = 6, family = "poisson", sd.errors = F,
    maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
while (is.infinite(logLik(mod4d))) {
    mod4d <- try(gllvm(y1, X = X, num.RR = 6, family = "poisson", sd.errors = F,
        maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
```

```
}
mod5d <- try(gllvm(y2, X = X, num.RR = 6, family = "binomial", sd.errors = F,
    maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
while (is.infinite(logLik(mod5d))) {
    mod5d <- try(gllvm(y2, X = X, num.RR = 6, family = "binomial",
        sd.errors = F, maxit = 1e+05, max.iter = 1e+05, starting.val = "res"),
        silent = T)
}
mod6d <- try(gllvm(y3, X = X, num.RR = 6, family = "gaussian", sd.errors = F,
    maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
while (is.infinite(logLik(mod6d))) {
    mod6d <- try(gllvm(y3, X = X, num.RR = 6, family = "gaussian",
        sd.errors = F, maxit = 1e+05, max.iter = 1e+05, starting.val = "res"),
        silent = T)
}

mod4e <- try(gllvm(y1, X = X, num.RR = 7, family = "poisson", sd.errors = F,
    maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
while (is.infinite(logLik(mod4e))) {
    mod4e <- try(gllvm(y1, X = X, num.RR = 7, family = "poisson", sd.errors = F,
        maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
}
mod5e <- try(gllvm(y2, X = X, num.RR = 7, family = "binomial", sd.errors = F,
    maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
while (is.infinite(logLik(mod5e))) {
    mod5e <- try(gllvm(y2, X = X, num.RR = 7, family = "binomial",
        sd.errors = F, maxit = 1e+05, max.iter = 1e+05, starting.val = "res"),
        silent = T)
}
mod6e <- try(gllvm(y3, X = X, num.RR = 7, family = "gaussian", sd.errors = F,
    maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
while (is.infinite(logLik(mod6e))) {
```

```
    mod6e <- try(gllvm(y3, X = X, num.RR = 7, family = "gaussian",
        sd.errors = F, maxit = 1e+05, max.iter = 1e+05, starting.val = "res"),
        silent = T)
}

mod4f <- try(gllvm(y1, X = X, num.RR = 8, family = "poisson", sd.errors = F,
    maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
while (is.infinite(logLik(mod4f))) {
    mod4f <- try(gllvm(y1, X = X, num.RR = 8, family = "poisson", sd.errors = F,
        maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
}
mod5f <- try(gllvm(y2, X = X, num.RR = 8, family = "binomial", sd.errors = F,
    maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
while (is.infinite(logLik(mod5f))) {
    mod5f <- try(gllvm(y2, X = X, num.RR = 8, family = "binomial",
        sd.errors = F, maxit = 1e+05, max.iter = 1e+05, starting.val = "res"),
        silent = T)
}
mod6f <- try(gllvm(y3, X = X, num.RR = 8, family = "gaussian", sd.errors = F,
    maxit = 1e+05, max.iter = 1e+05, starting.val = "res"), silent = T)
while (is.infinite(logLik(mod6f))) {
    mod6f <- try(gllvm(y3, X = X, num.RR = 8, family = "gaussian",
        sd.errors = F, maxit = 1e+05, max.iter = 1e+05, starting.val = "res"),
        silent = T)
}

# Put all the results together
result <- matrix(0, ncol = 3, nrow = 24)
colnames(result) <- c("Method", "Distribution", "RMSE")
result[, 1] <- c("GLM", "GLM", "LM", paste("RRR", rep(2:8, each = 3),
    sep = ""))
result[, 2] <- rep(c("Poisson", "Bernoulli", "Gaussian"), times = 8)
```

```
result[, 3] <- c(vegan::procrustes(mod1$coefficients[-1, ], mod$coefficients[-1,
    ], symmetric = TRUE)$ss, vegan::procrustes(mod2$coefficients[-1,
    ], mod1$coefficients[-1, ], symmetric = TRUE)$ss, vegan::procrustes(mod3$coefficients[-1,
    ], mod$coefficients[-1, ], symmetric = TRUE)$ss, vegan::procrustes(mod4$params$LvXcoef %*%
    t(mod4$params$theta), mod$coefficients[-1, ], symmetric = TRUE)$ss,
    vegan::procrustes(mod5$params$LvXcoef %*% t(mod5$params$theta),
        mod$coefficients[-1, ], symmetric = TRUE)$ss, vegan::procrustes(mod6$params$LvXcoef %*%
        t(mod6$params$theta), mod$coefficients[-1, ], symmetric = TRUE)$ss,
    vegan::procrustes(mod4a$params$LvXcoef %*% t(mod4a$params$theta),
        mod$coefficients[-1, ], symmetric = TRUE)$ss, vegan::procrustes(mod5a$params$LvXcoef %*%
        t(mod5a$params$theta), mod$coefficients[-1, ], symmetric = TRUE)$ss,
    vegan::procrustes(mod6a$params$LvXcoef %*% t(mod6a$params$theta),
        mod$coefficients[-1, ], symmetric = TRUE)$ss, vegan::procrustes(mod4b$params$LvXcoef %*%
        t(mod4b$params$theta), mod$coefficients[-1, ], symmetric = TRUE)$ss,
    vegan::procrustes(mod5b$params$LvXcoef %*% t(mod5b$params$theta),
        mod$coefficients[-1, ], symmetric = TRUE)$ss, vegan::procrustes(mod6b$params$LvXcoef %*%
        t(mod6b$params$theta), mod$coefficients[-1, ], symmetric = TRUE)$ss,
    vegan::procrustes(mod4c$params$LvXcoef %*% t(mod4c$params$theta),
        mod$coefficients[-1, ], symmetric = TRUE)$ss, vegan::procrustes(mod5c$params$LvXcoef %*%
        t(mod5c$params$theta), mod$coefficients[-1, ], symmetric = TRUE)$ss,
    vegan::procrustes(mod6c$params$LvXcoef %*% t(mod6c$params$theta),
        mod$coefficients[-1, ], symmetric = TRUE)$ss, vegan::procrustes(mod4d$params$LvXcoef %*%
        t(mod4d$params$theta), mod$coefficients[-1, ], symmetric = TRUE)$ss,
    vegan::procrustes(mod5d$params$LvXcoef %*% t(mod5d$params$theta),
        mod$coefficients[-1, ], symmetric = TRUE)$ss, vegan::procrustes(mod6d$params$LvXcoef %*%
        t(mod6d$params$theta), mod$coefficients[-1, ], symmetric = TRUE)$ss,
    vegan::procrustes(mod4e$params$LvXcoef %*% t(mod4e$params$theta),
        mod$coefficients[-1, ], symmetric = TRUE)$ss, vegan::procrustes(mod5e$params$LvXcoef %*%
        t(mod5e$params$theta), mod$coefficients[-1, ], symmetric = TRUE)$ss,
    vegan::procrustes(mod6e$params$LvXcoef %*% t(mod6e$params$theta),
        mod$coefficients[-1, ], symmetric = TRUE)$ss, vegan::procrustes(mod4f$params$LvXcoef %*%
        t(mod4f$params$theta), mod$coefficients[-1, ], symmetric = TRUE)$ss,
```

```
        vegan::procrustes(mod5f$params$LvXcoef %*% t(mod5f$params$theta),
            mod$coefficients[-1, ], symmetric = TRUE)$ss, vegan::procrustes(mod6f$params$LvXcoef %*%
            t(mod6f$params$theta), mod$coefficients[-1, ], symmetric = TRUE)$ss)
    print(r)
    totres <- rbind(totres, cbind(result, sim = r))
}
```

Code for the second simulation study

```
library(gllvm)
library(vegan)
R <- 1000  #number of simulations
n <- 100  #number of sites
p <- 30  #number of species
num.lv <- 2  #number of latent variables
set.seed(1)
beta0 <- matrix(runif(p, -1, 1), ncol = p, nrow = n, byrow = T)
set.seed(1)
num.X <- 5
X <- mvtnorm::rmvnorm(n, rep(0, num.X), diag(num.X))
colnames(X) <- 1:num.X
set.seed(1)
epsilon <- mvtnorm::rmvnorm(n, rep(0, num.lv), diag(num.lv))
epsilon <- resid(lm(epsilon ~ X))  #make sure that the residual is independent from the covariates as w
set.seed(1)
gamma <- matrix(runif(p * num.lv, -2, 2), ncol = num.lv)
gamma[upper.tri(gamma)] <- 0
diag(gamma) <- 1
set.seed(1)
beta <- matrix(factanal(X, factors = 2)$loadings, ncol = 2)
LV <- (X %*% (beta) + epsilon)
eta <- beta0 + LV %*% t(gamma)
```

```r
eta2 <- beta0 + epsilon %*% t(gamma)


totresult <- NULL
for (r in 1:R) {
    set.seed(r)
    y <- matrix(rpois(p * n, exp(eta)), ncol = p, nrow = n)
    set.seed(r)
    y2 <- matrix(rpois(p * n, exp(eta2)), ncol = p, nrow = n)
    set.seed(r)
    y3 <- matrix(rbinom(p * n, size = 1, prob = pnorm(eta)), ncol = p,
        nrow = n)
    set.seed(r)
    y4 <- matrix(rbinom(p * n, size = 1, prob = pnorm(eta2)), ncol = p,
        nrow = n)
    set.seed(r)
    y5 <- matrix(rnorm(p * n, mean = eta), ncol = p, nrow = n)
    set.seed(r)
    y6 <- matrix(rnorm(p * n, mean = eta2), ncol = p, nrow = n)

    # only fixed-effects
    set.seed(r)
    y7 <- matrix(rpois(p * n, exp(eta3)), ncol = p, nrow = n)
    set.seed(r)
    y8 <- matrix(rbinom(p * n, size = 1, prob = pnorm(eta3)), ncol = p,
        nrow = n)
    set.seed(r)
    y9 <- matrix(rnorm(p * n, mean = eta3), ncol = p, nrow = n)

    LV1 <- LV
    LV2 <- epsilon
    LV3 <- LV
    LV4 <- epsilon
```

```
LV5 <- LV
LV6 <- epsilon
LV7 <- LVb
LV8 <- LVb
LV9 <- LVb

# remove empty rows here for CA and CCA...
idx1 <- rowSums(y)
idx2 <- rowSums(y2)
idx3 <- rowSums(y3)
idx4 <- rowSums(y4)
idx7 <- rowSums(y7)
idx8 <- rowSums(y8)

X1 <- X[idx1 > 0, ]
X2 <- X[idx2 > 0, ]
X3 <- X[idx3 > 0, ]
X4 <- X[idx4 > 0, ]
X7 <- X[idx7 > 0, ]
X8 <- X[idx8 > 0, ]

y <- y[idx1 > 0, ]
LV1 <- LV1[idx1 > 0, ]
y2 <- y2[idx2 > 0, ]
LV2 <- LV2[idx2 > 0, ]
y3 <- y3[idx3 > 0, ]
LV3 <- LV3[idx3 > 0, ]
y4 <- y4[idx4 > 0, ]
LV4 <- LV4[idx4 > 0, ]
y7 <- y7[idx7 > 0, ]
LV7 <- LV7[idx7 > 0, ]
y8 <- y8[idx8 > 0, ]
```

```
LV8 <- LV8[idx8 > 0, ]

# Run the models. Loop to ensure models don't converge to infinity

# Poisson Constrained simulation
modPC <- try(gllvm(y = y, X = X1, num.lv.c = num.lv, family = "poisson",
    starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
    silent = T)
if (inherits(modPC, "try-error")) {
    modPC <- try(gllvm(y = y, X = X1, num.lv.c = num.lv, family = "poisson",
        starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
        silent = T)
    while (inherits(modPC, "try-error")) {
        modPC <- try(gllvm(y = y, X = X1, num.lv.c = num.lv, family = "poisson",
            starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
            silent = T)
    }
}
PCCA <- cca(y, X1)
# Unconstrained simulation
modPCa <- try(gllvm(y = y2, X = X2, num.lv.c = num.lv, family = "poisson",
    starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
    silent = T)
if (inherits(modPCa, "try-error")) {
    modPCa <- try(gllvm(y = y2, X = X2, num.lv.c = num.lv, family = "poisson",
        starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
        silent = T)
    while (inherits(modPCa, "try-error")) {
        modPCa <- try(gllvm(y = y2, X = X2, num.lv.c = num.lv, family = "poisson",
            starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
            silent = T)
    }
```

```
}

PCCAa <- cca(y2, X2)

# Fixed-effects simulation
modPCb <- try(gllvm(y = y7, X = X7, num.RR = num.lv, family = "poisson",
    starting.val = "zero", maxit = 1e+07, num.lv = 0, sd.errors = F),
    silent = T)
if (inherits(modPCb, "try-error")) {
    modPCb <- try(gllvm(y = y7, X = X7, num.RR = num.lv, family = "poisson",
        starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
        silent = T)
    while (inherits(modPCb, "try-error")) {
        modPCb <- try(gllvm(y = y7, X = X7, num.RR = num.lv, family = "poisson",
            starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
            silent = T)
    }
}

PCCAb <- cca(y7, X7)

# Bernoulli
modBC <- try(gllvm(y = y3, X = X3, num.lv.c = num.lv, family = "binomial",
    starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
    silent = T)

if (inherits(modBC, "try-error")) {
    modBC <- try(gllvm(y = y3, X = X3, num.lv.c = num.lv, family = "binomial",
        starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
        silent = T)
    while (inherits(modBC, "try-error")) {
        modBC <- try(gllvm(y = y3, X = X3, num.lv.c = num.lv, family = "binomial",
```

```
            starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
            silent = T)
    }
} else if (is.infinite(logLik(modBC))) {
    modBC <- try(gllvm(y = y3, X = X3, num.lv.c = num.lv, family = "binomial",
        starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
        silent = T)
    while (inherits(modBC, "try-error")) {
        modBC <- try(gllvm(y = y3, X = X3, num.lv.c = num.lv, family = "binomial",
            starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
            silent = T)
    }
    while (is.infinite(logLik(modBC))) {
        modBC <- try(gllvm(y = y3, X = X3, num.lv.c = num.lv, family = "binomial",
            starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
            silent = T)
    }
}

BCCA <- cca(y3, X3)
# Unconstrained simulation
modBCa <- try(gllvm(y = y4, X = X4, num.lv.c = num.lv, family = "binomial",
    starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
    silent = T)
if (inherits(modBCa, "try-error")) {
    modBCa <- try(gllvm(y = y4, X = X4, num.lv.c = num.lv, family = "binomial",
        starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
        silent = T)
    while (inherits(modBCa, "try-error")) {
        modBCa <- try(gllvm(y = y4, X = X4, num.lv.c = num.lv, family = "binomial",
            starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
            silent = T)
```

```
    }
} else if (is.infinite(logLik(modBCa))) {
    modBCa <- try(gllvm(y = y4, X = X4, num.lv.c = num.lv, family = "binomial",
        starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
        silent = T)
    while (inherits(modBCa, "try-error")) {
        modBCa <- try(gllvm(y = y4, X = X4, num.lv.c = num.lv, family = "binomial",
            starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
            silent = T)
    }
    while (is.infinite(logLik(modBCa))) {
        modBCa <- try(gllvm(y = y4, X = X4, num.lv.c = num.lv, family = "binomial",
            starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
            silent = T)
    }
}
BCCAa <- cca(y4, X4)

# fixed-effects simulation
modBCb <- try(gllvm(y = y8, X = X8, num.RR = num.lv, family = "binomial",
    starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
    silent = T)
if (inherits(modBCb, "try-error")) {
    modBCb <- try(gllvm(y = y8, X = X8, num.RR = num.lv, family = "binomial",
        starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
        silent = T)
    while (inherits(modBCb, "try-error")) {
        modBCb <- try(gllvm(y = y8, X = X8, num.RR = num.lv, family = "binomial",
            starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
            silent = T)
    }
} else if (is.infinite(logLik(modBCb))) {
```

```
    modBCb <- try(gllvm(y = y8, X = X8, num.RR = num.lv, family = "binomial",
        starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
        silent = T)
    while (inherits(modBCb, "try-error")) {
        modBCb <- try(gllvm(y = y8, X = X8, num.RR = num.lv, family = "binomial",
            starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
            silent = T)
    }
    while (is.infinite(logLik(modBCb))) {
        modBCb <- try(gllvm(y = y8, X = X8, num.RR = num.lv, family = "binomial",
            starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
            silent = T)
    }
}
BCCAb <- cca(y8, X8)

# Constrained simulation
modGC <- try(gllvm(y = y5, X = X, num.lv.c = num.lv, family = "gaussian",
    starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
    silent = T)
if (inherits(modGC, "try-error")) {
    modGC <- try(gllvm(y = y5, X = X, num.lv.c = num.lv, family = "gaussian",
        starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
        silent = T)
    while (inherits(modGC, "try-error")) {
        modGC <- try(gllvm(y = y5, X = X, num.lv.c = num.lv, family = "gaussian",
            starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
            silent = T)
    }
} else if (is.infinite(logLik(modGC))) {
    modGC <- try(gllvm(y = y5, X = X, num.lv.c = num.lv, family = "gaussian",
        starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
```

```
        silent = T)
    while (inherits(modGC, "try-error")) {
        modGC <- try(gllvm(y = y5, X = X, num.lv.c = num.lv, family = "gaussian",
            starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
            silent = T)
    }
    while (is.infinite(logLik(modGC))) {
        modGC <- try(gllvm(y = y5, X = X, num.lv.c = num.lv, family = "gaussian",
            starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
            silent = T)
    }
}

RDA <- rda(y5, X)
# Unconstrained simulation
modGCa <- try(gllvm(y = y6, X = X, num.lv.c = num.lv, family = "gaussian",
    starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
    silent = T)
if (inherits(modGCa, "try-error")) {
    modGCa <- try(gllvm(y = y6, X = X, num.lv.c = num.lv, family = "gaussian",
        starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
        silent = T)
    while (inherits(modGCa, "try-error")) {
        modGCa <- try(gllvm(y = y6, X = X, num.lv.c = num.lv, family = "gaussian",
            starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
            silent = T)
    }
} else if (is.infinite(logLik(modGCa))) {
    modGCa <- try(gllvm(y = y6, X = X, num.lv.c = num.lv, family = "gaussian",
        starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
        silent = T)
    while (inherits(modGCa, "try-error")) {
```

```
        modGCa <- try(gllvm(y = y6, X = X, num.lv.c = num.lv, family = "gaussian",
            starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
            silent = T)
    }
    while (is.infinite(logLik(modGCa))) {
        modGCa <- try(gllvm(y = y6, X = X, num.lv.c = num.lv, family = "gaussian",
            starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
            silent = T)
    }
}

RDAa <- rda(y6, X)

# fixed-effects simulation
modGCb <- try(gllvm(y = y9, X = X, num.RR = num.lv, family = "gaussian",
    starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
    silent = T)
if (inherits(modGCb, "try-error")) {
    modGCb <- try(gllvm(y = y9, X = X, num.RR = num.lv, family = "gaussian",
        starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
        silent = T)
    while (inherits(modGCb, "try-error")) {
        modGCb <- try(gllvm(y = y9, X = X, num.RR = num.lv, family = "gaussian",
            starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
            silent = T)
    }
} else if (is.infinite(logLik(modGCb))) {
    modGCb <- try(gllvm(y = y9, X = X, num.RR = num.lv, family = "gaussian",
        starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
        silent = T)
    while (inherits(modGCb, "try-error")) {
        modGCb <- try(gllvm(y = y9, X = X, num.RR = num.lv, family = "gaussian",
```

```
            starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
            silent = T)
    }
    while (is.infinite(logLik(modGCb))) {
        modGCb <- try(gllvm(y = y9, X = X, num.RR = num.lv, family = "gaussian",
            starting.val = "res", maxit = 1e+07, num.lv = 0, sd.errors = F),
            silent = T)
    }
}

RDAb <- rda(y9, X)

result <- matrix(0, ncol = 5, nrow = 18)
colnames(result) <- c("Simulation", "Method", "Distribution", "SS",
    "SS_spp")
result[, 1] <- rep(c("Constrained with residual", "Unconstrained",
    "Constrained without residual"), times = 6)
result[, 2] <- c(rep(c("GLLVM", "CCA"), each = 6), rep(c("GLLVM", "RDA"),
    each = 3))
result[, 3] <- c(rep(c(rep("Poisson", 3), rep("Bernoulli", 3)), times = 2),
    rep("Gaussian", 6))
result[, 4] <- c(procrustes(getLV(modPC), LV1, symmetric = T)$ss, procrustes(getLV(modPCa),
    LV2, symmetric = T)$ss, procrustes(getLV(modPCb), LV7, symmetric = T)$ss,
    procrustes(getLV(modBC), LV3, symmetric = T)$ss, procrustes(getLV(modBCa),
        LV4, symmetric = T)$ss, procrustes(getLV(modBCb), LV8, symmetric = T)$ss,
    procrustes(scores(PCCA)$sites, LV1, symmetric = T)$ss, procrustes(scores(PCCAa)$sites,
        LV2, symmetric = T)$ss, procrustes(scores(PCCAb)$sites, LV7,
        symmetric = T)$ss, procrustes(scores(BCCA)$sites, LV3, symmetric = T)$ss,
    procrustes(scores(BCCAa)$sites, LV4, symmetric = T)$ss, procrustes(scores(BCCAb)$sites,
        LV8, symmetric = T)$ss, procrustes(getLV(modGC), LV5, symmetric = T)$ss,
    procrustes(getLV(modGCa), LV6, symmetric = T)$ss, procrustes(getLV(modGCb),
        LV9, symmetric = T)$ss, procrustes(scores(RDA)$sites, LV5,
```

```
            symmetric = T)$ss, procrustes(scores(RDAa)$sites, LV6, symmetric = T)$ss,
        procrustes(scores(RDAb)$sites, LV9, symmetric = T)$ss)
    result[, 5] <- c(procrustes(modPC$params$theta, gamma, symmetric = T)$ss,
        procrustes(modPCa$params$theta, gamma, symmetric = T)$ss, procrustes(modPCb$params$theta,
            gamma, symmetric = T)$ss, procrustes(modBC$params$theta, gamma,
            symmetric = T)$ss, procrustes(modBCa$params$theta, gamma, symmetric = T)$ss,
        procrustes(modBCb$params$theta, gamma, symmetric = T)$ss, procrustes(scores(PCCA)$species,
            gamma, symmetric = T)$ss, procrustes(scores(PCCAa)$species,
            gamma, symmetric = T)$ss, procrustes(scores(PCCAb)$species,
            gamma, symmetric = T)$ss, procrustes(scores(BCCA)$species,
            gamma, symmetric = T)$ss, procrustes(scores(BCCAa)$species,
            gamma, symmetric = T)$ss, procrustes(scores(BCCAb)$species,
            gamma, symmetric = T)$ss, procrustes(modGC$params$theta, gamma,
            symmetric = T)$ss, procrustes(modGC$params$theta, gamma, symmetric = T)$ss,
        procrustes(modGCb$params$theta, gamma, symmetric = T)$ss, procrustes(scores(RDA)$species,
            gamma, symmetric = T)$ss, procrustes(scores(RDAa)$species,
            gamma, symmetric = T)$ss, procrustes(scores(RDAb)$species,
            gamma, symmetric = T)$ss)
    totresult <- rbind(totresult, cbind(result, sim = r))
    print(r)
}
```

Appendix S7: Extra figures and tables for examples

Table S1: Estimates, standard errors, Wald-statistic and accompanying p-values for predictors in the constrained ordination of the Swiss alpine plants data, rotated to principal direction. DDGE0 = Growing degree days above zero; MIND = Moisture index; SOLRAD = Total solar radiation over the year; TPI = Topography index.

	Estimate	Std. Error	z value	Pr(>|z|)
LV1				
DDEG0	1.126	0.238	4.742	0.000
Slope	-0.601	0.062	-9.732	0.000
MIND	0.452	0.083	5.463	0.000
SOLRAD	-0.684	0.070	-9.835	0.000
TPI	-0.133	0.030	-4.415	0.000
Elevation	4.745	0.428	11.085	0.000
LV2				
DDEG0	0.858	0.093	9.204	0.000
Slope	0.238	0.016	15.379	0.000
MIND	0.057	0.040	1.429	0.153
SOLRAD	0.109	0.019	5.598	0.000
TPI	-0.074	0.014	-5.246	0.000
Elevation	0.801	0.096	8.373	0.000

Figure S1: CCA for the Swiss alpine plants data

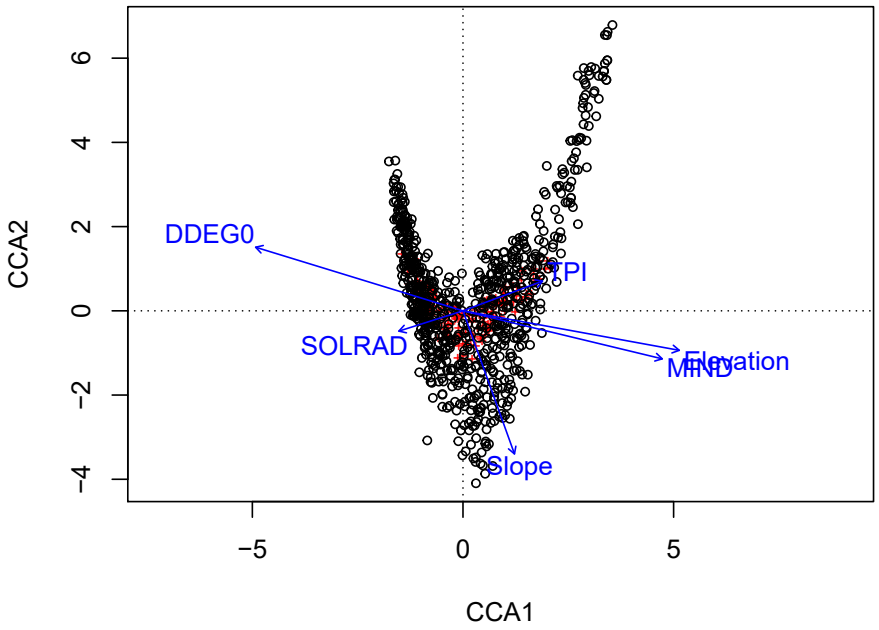


Table S2: List of species names for the Swiss alpine plants data

Species name	abbreviation
Achillea atrata	Acat
Achillea millefolium	Acmi
Acinos alpinus	Acal
Agrostis capillaris	Agca
Agrostis rupestris	Agru
Agrostis schraderiana	Agsc
Agrostis stolonifera	Agst
Ajuga reptans	Ajre
Alchemilla conjuncta	Alco

(continued)

Species name	abbreviation
Alchemilla glabra	Algl
Alchemilla vulgaris	Alvu
Androsace chamaejasme	Anch
Anthoxanthum odoratum	Anod
Anthyllis vulneraria	Anvu
Aposeris foetida	Apfo
Arabis alpina	Aral
Arnica montana	Armo
Arrhenatherum elatius	Arel
Aster bellidiastrum	Asbe
Astrantia major	Asma
Bartsia alpina	Baal
Bellis perennis	Bepe
Brachypodium pinnatum	Brpi
Briza media	Brme
Bromus erectus	Brer
Calamagrostis varia	Cava
Campanula barbata	Caba
Campanula scheuchzeri	Casc
Carduus defloratus	Cade
Carex atrata	Caat
Carex ferruginea	Cafe
Carex flacca	Cafl
Carex pallescens	Capa
Carex sempervirens	Case
Carex sylvatica	Casy
Carlina acaulis	Caac
Carum carvi	Caca

(continued)

Species name	abbreviation
Centaurea jacea	Ceja
Centaurea montana	Cemo
Centaurea scabiosa	Cesc
Cerastium arvense	Cear
Cerastium fontanum	Cefo
Chaerophyllum hirsutum	Chhi
Cirsium acaule	Ciac
Cirsium spinosissimum	Cisp
Clinopodium vulgare	Clvu
Crepis aurea	Crau
Crepis pyrenaica	Crpy
Cruciata laevipes	Crla
Cynosurus cristatus	Cycr
Dactylis glomerata	Dagl
Daucus carota	Daca
Deschampsia cespitosa	Dece
Doronicum grandiflorum	Dogr
Dryas octopetala	Droc
Euphorbia cyparissias	Eucy
Festuca pratensis	Fepr
Festuca quadriflora	Fequ
Festuca rubra	Feru
Festuca violacea	Fevi
Fragaria vesca	Frve
Galium album	Gaal
Galium anisophyllon	Gaan
Galium pumilum	Gapu
Gentiana acaulis	Geac

(continued)

Species name	abbreviation
Gentiana campestris	Geca
Gentiana lutea	Gelu
Gentiana purpurea	Gepu
Gentiana verna	Geve
Geranium sylvaticum	Gesy
Geum montanum	Gemo
Geum rivale	Geri
Glechoma hederacea	Glhe
Globularia cordifolia	Glco
Globularia nudicaulis	Glnu
Hedysarum hedysaroides	Hehe
Helianthemum nummularium	Henu
Helictotrichon versicolor	Heve
Heracleum sphondylium	Hesp
Hieracium bifidum	Hibi
Hieracium lactucella	Hila
Hippocrepis comosa	Hico
Holcus lanatus	Hola
Homogyne alpina	Hoal
Hypericum maculatum	Hyma
Hypochaeris radicata	Hyra
Knautia arvensis	Knar
Knautia dipsacifolia	Kndi
Laserpitium latifolium	Lala
Lathyrus pratensis	Lapr
Leontodon autumnalis	Leau
Leontodon helveticus	Lehe
Leontodon hispidus	Lehi

(continued)

Species name	abbreviation
Leucanthemum vulgare	Levu
Ligusticum mutellina	Limu
Linum catharticum	Lica
Lolium perenne	Lope
Lotus corniculatus	Loco
Luzula multiflora	Lumu
Medicago lupulina	Melu
Myosotis alpestris	Myal
Nardus stricta	Nast
Parnassia palustris	Papa
Pedicularis foliosa	Pefo
Phleum hirsutum	Phhi
Phleum pratense	Phpr
Phleum rhaeticum	Phrh
Phyteuma orbiculare	Phor
Phyteuma spicatum	Phsp
Pimpinella major	Pima
Plantago alpina	Plal
Plantago atrata	Plat
Plantago lanceolata	Plla
Plantago major	Plma
Plantago media	Plme
Poa alpina	Poal
Poa minor	Pomi
Poa pratensis	Popr
Polygala chamaebuxus	Poch
Polygonum bistorta	Pobi
Polygonum viviparum	Povi

(continued)

Species name	abbreviation
Potentilla aurea	Poau
Potentilla crantzii	Pocr
Potentilla erecta	Poer
Primula elatior	Prel
Pritzelago alpina	Pral
Prunella grandiflora	Prgr
Prunella vulgaris	Prvu
Pulsatilla alpina	Pual
Ranunculus aconitifolius	Raac
Ranunculus acris	Raac
Ranunculus alpestris	Raal
Ranunculus montanus	Ramo
Ranunculus repens	Rare
Rhinanthus alectorolophus	Rhal
Rhinanthus minor	Rhmi
Rumex acetosa	Ruac
Salix herbacea	Sahe
Salix reticulata	Sare
Salix retusa	Sare
Sanguisorba minor	Sami
Saxifraga aizoides	Saai
Saxifraga moschata	Samo
Saxifraga oppositifolia	Saop
Saxifraga paniculata	Sapa
Scabiosa lucida	Sclu
Sesleria caerulea	Seca
Silene acaulis	Siac
Silene vulgaris	Sivu
Soldanella alpina	Soal

(continued)

Species name	abbreviation
Solidago virgaurea	Sovi
Stachys officinalis	Stof
Taraxacum alpinum	Taal
Taraxacum officinale	Taof
Thesium alpinum	Thal
Thymus praecox	Thpr
Tragopogon pratensis	Trpr
Trifolium badium	Trba
Trifolium medium	Trme
Trifolium pratense	Trpr
Trifolium repens	Trre
Trifolium thalii	Trth
Trisetum flavescens	Trfl
Trollius europaeus	Treu
Vaccinium gaultherioides	Vaga
Vaccinium myrtillus	Vamy
Vaccinium vitisidaea	Vavi
Valeriana montana	Vamo
Veratrum album	Veal
Veronica alpina	Veal
Veronica chamaedrys	Vech
Veronica officinalis	Veof
Vicia cracca	Vicr
Vicia sepium	Vise
Viola biflora	Vibi

Figure S2: Partial CCA for the Levanger grasslands data

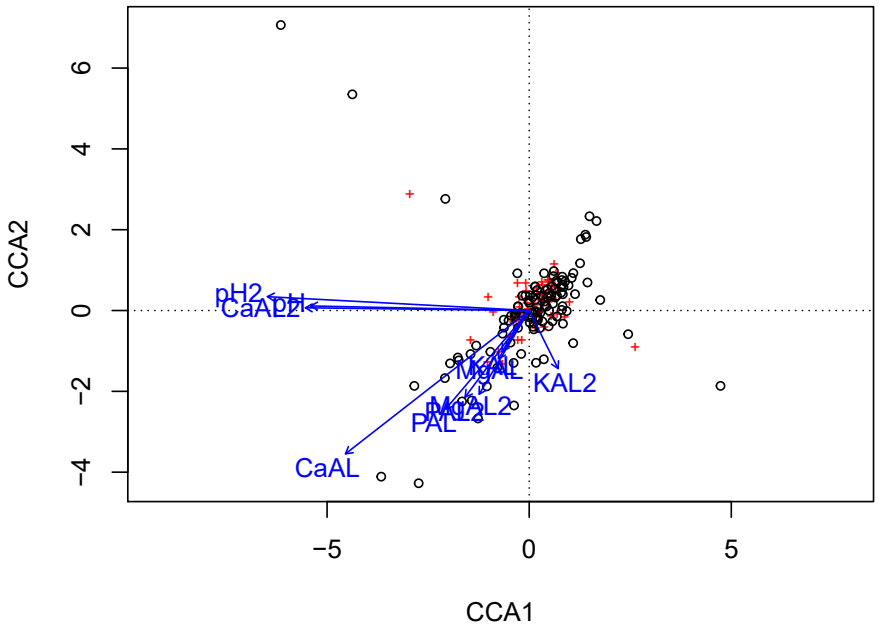


Table S3: Estimates, standard errors, Wald-statistic and accompanying p-values for predictors in the constrained ordination of the Levanger grasslands data, including various predictors of water-soluble soil nutrients for 0 - 10cm and 10 - 20cm soil depth, rotated to principal direction.

	Estimate	Std. Error	z value	Pr(>|z|)
LV1				
pH (0 - 10)	0.671	0.409	1.640	0.101
pH (10 - 20)	-0.241	0.325	-0.743	0.458
Phosphorus (0 - 10)	-0.380	0.337	-1.127	0.260
Phosphorus (10 - 20)	0.070	0.310	0.224	0.822
Potassium (0 - 10)	0.049	0.195	0.251	0.802
Potassium (10 - 20)	-0.302	0.187	-1.612	0.107
Magnesium (0 - 10)	0.512	0.295	1.735	0.083
Magnesium (10 - 20)	-0.218	0.265	-0.824	0.410
Calcium (0 - 10)	-1.034	0.358	-2.889	0.004
Calcium (10 - 20)	0.028	0.268	0.105	0.916
LV2				
pH (0 - 10)	0.556	0.278	2.000	0.045
pH (10 - 20)	0.445	0.314	1.418	0.156
Phosphorus (0 - 10)	0.612	0.412	1.486	0.137
Phosphorus (10 - 20)	-0.193	0.282	-0.685	0.493
Potassium (0 - 10)	0.260	0.158	1.652	0.098
Potassium (10 - 20)	-0.147	0.155	-0.947	0.344
Magnesium (0 - 10)	-0.640	0.413	-1.548	0.122
Magnesium (10 - 20)	0.525	0.295	1.777	0.076
Calcium (0 - 10)	-0.148	0.343	-0.431	0.666
Calcium (10 - 20)	-0.310	0.226	-1.376	0.169

Table S4: List of species names for the Levanger grasslands data

Species name	abbreviation
Urtica dioica	Urdi

(continued)

Species name	abbreviation
Rumex longifolius	Rulo
Rumex acetosa	Ruac
Bistorta vivipara	Bivi
Stellaria media	Stme
Stellaria graminea	Stgr
Cerastium fontanum	Cefo
Ranunculus auricomus	Raau
Ranunculus acris	Raac
Ranunculus repens	Rare
Anemone nemorosa	Anne
Filipendula ulmaria	Fiul
Geum urbanum	Geur
Potentilla erecta	Poer
Fragaria vesca	Frve
Alchemilla sp	Alsp
Trifolium repens	Trre
Trifolium pratense	Trpr
Lotus corniculatus	Loco
Vicia cracca	Vicr
Lathyrus pratensis	Lapr
Oxalis acetosella	Oxac
Geranium sylvaticum	Gesy
Hypericum maculatum	Hyma
Viola riviniana	Viri
Viola canina	Vica
Anthriscus sylvestris	Ansy
Carum carvi	Caca
Pimpinella saxifraga	Pisa
Heracleum sp	Hesp

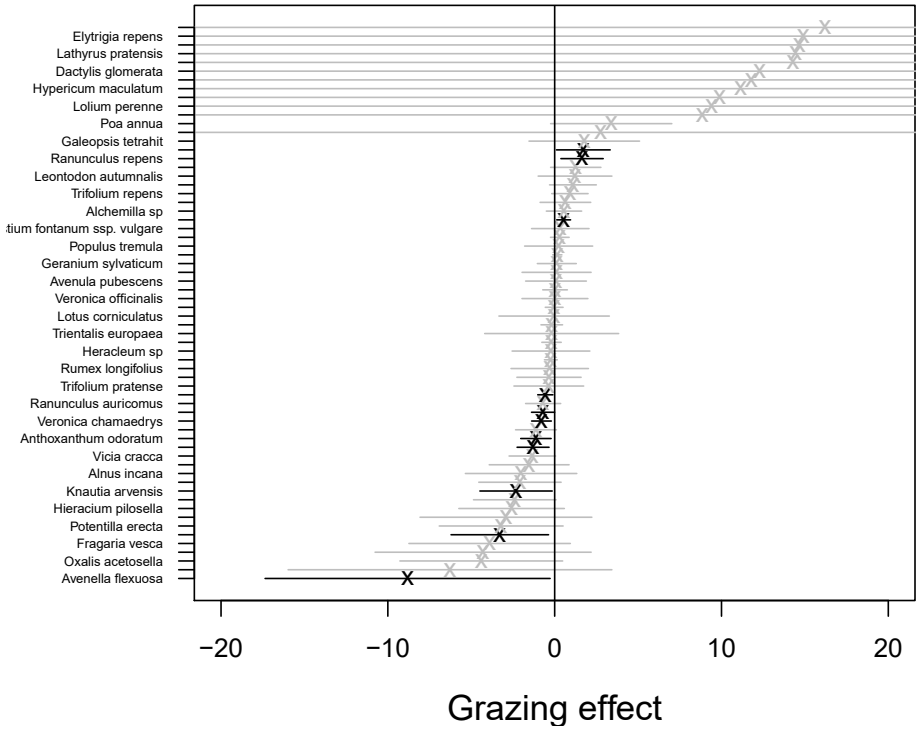
(continued)

Species name	abbreviation
Trientalis europaea	Treu
Galium boreale	Gabo
Galeopsis tetrahit	Gate
Veronica chamaedrys	Vech
Veronica serpyllifolia	Vese
Veronica officinalis	Veof
Knautia arvensis	Knar
Campanula rotundifolia	Caro
Achillea millefolium	Acmi
Achillea ptarmica	Acpt
Leontodon autumnalis	Leau
Taraxacum sp	Tasp
Hieracium sp	Hisp
Hieracium pilosella	Hipi
Maianthemum bifolium	Mabi
Luzula pilosa	Lupi
Carex pallescens	Capa
Anthoxanthum odoratum	Anod
Phleum pratense	Phpr
Agrostis capillaris	Agca
Avenella flexuosa	Avfl
Deschampsia cespitosa	Dece
Avenula pubescens	Avpu
Dactylis glomerata	Dagl
Poa pratensis	Popr
Poa annua	Poan
Poa trivialis	Potr
Festuca rubra	Feru

(continued)

Species name	abbreviation
Schedonorus pratensis	Scpr
Lolium perenne	Lope
Elytrigia repens	Elre
Populus tremula	Potr
Sorbus aucuparia	Soau
Alnus incana	Alin

Figure S3: Effects of grazing for the Levanger grasslands data



References

Austin, M. (2007). Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling*, **200**, 1–19. Retrieved October 13, 2021, from https://www.sciencedirect.com/science/article/pii/S0304380006003140

Booth, J.G. & Hobert, J.P. (1998). Standard Errors of Prediction in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, **93**, 262–272. Retrieved June 15, 2021, from https://www.tandfonline.com/doi/abs/10.1080/01621459.1998.10474107

Hui, F.K.C., Warton, D.I., Ormerod, J.T., Haapaniemi, V. & Taskinen, S. (2017). Variational Approximations for Generalized Linear Latent Variable Models. *Journal of Computational and Graphical Statistics*, **26**, 35–43. Retrieved April 24, 2020, from https://doi.org/10.1080/10618600.2016.1164708

Niku, J. (2020). On modeling multivariate abundance data with generalized linear latent variable models. *JYU dissertations*. Retrieved June 30, 2021, from https://jyx.jyu.fi/handle/123456789/67735

Niku, J., Warton, D.I., Hui, F.K.C. & Taskinen, S. (2017). Generalized Linear Latent Variable Models for Multivariate Count and Biomass Data in Ecology. *JABES*, **22**, 498–522. Retrieved April 24, 2020, from https://doi.org/10.1007/s13253-017-0304-7

ter Braak, C.J. (1986). Canonical Correspondence Analysis: A New Eigenvector Technique for Multivariate Direct Gradient Analysis. *Ecology*, **67**, 1167–1179. Retrieved May 25, 2020, from https://esajournals.onlinelibrary.wiley.com/doi/abs/10.2307/1938672

van der Veen, B., Hui, F.K.C., Hovstad, K.A., Solbu, E.B. & O'Hara, R.B. (2021). Model-based ordination for species with unequal niche widths. *Methods in Ecology and Evolution*, **n/a**. Retrieved April 20, 2021, from https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13595

Yee, T.W. (2004). A New Technique for Maximum-Likelihood Canonical Gaussian Ordination. *Ecological Monographs*, **74**, 685–701. Retrieved June 9, 2020, from https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/03-0078

Yee, T.W. (2015). Constrained Quadratic Ordination. *Vector Generalized Linear and Additive Models: With an Implementation in R* (ed T.W. Yee), pp. 201–237. Springer Series in Statistics. Springer, New York, NY. Retrieved October 13, 2021, from https://doi.org/10.1007/978-1-4939-2818-7_6

1.3 Model-based analysis of niche overlap with Generalized Linear Latent Variable Models

Model-based analysis of niche overlap with Generalized Linear Latent Variable Models

Bert van der Veen[123] Robert B. O'Hara[23] Francis K.C. Hui[4]
Knut A. Hovstad[35]

[1]Department of Landscape and Biodiversity, Norwegian Institute of Bioeconomy research, Trondheim, Norway
[2]Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway
[3]Centre of Biodiversity Dynamics, Norwegian University of Science and Technology, Trondheim, Norway
[4]Research School of Finance, Actuarial Studies and Statistics, The Australian National University, Canberra, Australia
[5]The Norwegian Biodiversity Information Centre, Trondheim, Norway

Summary

The ecological niche is a fundamental concept in ecology that can be used in order better understand species relationships. The overlap in species niches provides a measure of the likelihood for species to co-occur. Most approaches that quantify niche overlap have been based on distance and similarity indices, for pairwise combinations of species. In this article, we present a model-based approach for the prediction of niche overlap for commonly used regression methods, such as (joint) species distribution models. We do this using an example dataset of an ecological community of Foraminifera species, to which we fit a Generalized Linear Latent Variable Model (GLLVM). GLLVMs are a flexible class of models that allow to estimate the distribution of species using both measured and unmeasured components, with additional benefits such as access to statistical uncertainty of parameter estimates. We demonstrate how to calculate niche overlap from GLLVMs for any combination of species, and separately for different environments. Model-based analysis of niche overlap further expands the toolset available to ecologists for the exploration of species co-occurrence patterns. **keywords**: model-based ordination, unimodal

response, niche model

Introduction

The ecological niche of a species is a fundamental concept in ecology, describing the relationship between a species and the environment (Grinnell 1924; Elton 1927; Hutchinson 1959). Since popularisation of the niche concept (Hutchinson 1957), it has been eagerly utilized by community ecologists to infer on species relationships (MacArthur & Levins 1967) e.g., through calculations of niche overlap (May & Arthur 1972; Holt 1987).

Measures of niche overlap help ecologists to compare species' resource preferences (Wathne *et al.* 2000; Vogel *et al.* 2019) so they can understand why multiple species coexist (or not) even without competition. Niche overlap theory is inherently connected with species coexistence theory (Pianka 1974; Holt 1987; Chesson 2000; Letten *et al.* 2017) e.g., as the magnitude of overlap in the exploitation of limiting resources between co-occurring species is key to predicting their stable coexistence or local exclusion (Chase & Leibold 2003; Letten *et al.* 2017). According to the principle of competitive exclusion (Hardin 1960), two species competing for the same limited resource cannot coexist at constant population values (Gause 1934). Thus, species occur in environments where they are able to exploit resources in ways that render them competitively superior or able to coexist with others, and are absent from environments in which resources are insufficient (Godsoe & Harmon 2012).

A range of metrics have been developed to quantify niche overlap. Early approaches were mostly based on the calculation of distance and similarity indices, from abundance data (Pianka 1973; Hurlbert 1978). Some of these indices, for example Pianka's niche overlap index, require that the relative proportion of the resource in question used by each of the species can be quantified, which is often difficult or not possible in practice. To overcome this obstacle, distance of similarity in functional traits are often used as a proxy to examine and estimate niche overlap (McGill *et al.* 2006). Darwin's study of beak size and beak depth in coexisting finch species at the Galapagos Islands is a classic example of this (Darwin 1859). Similar distance-based approaches have been developed to estimate niche overlap in food webs (Cohen 1977; Cohen 1978; Cattin *et al.* 2004).

Most commonly the ecological niche of a species is modelled using a regression-based approach (Austin *et al.* 1990; Guisan & Zimmermann 2000; Jansen & Oksanen 2013; van der Veen *et al.* 2021b). The prediction of species niches using regression methods corresponds well with the Hutchinsonian niche concept (Hutchinson 1959), as the niche is represented in multiple dimensions by measured resources or the environment (e.g. with predictors). The shape of response curve, as represented using functions of predictors, represents a modellers'

expectation for the shape of the hypersurface, i.e. the shape of a species' niche (Austin 1987). Since the responses of species are generally unknown in a regression, maximum likelihood estimation can be used to retrieve parameter estimates, and to optimally represent species niches.

Examples of univariate methods that can be used to estimate species niches are Generalized Linear Models (GLMs, Nelder & Wedderburn 1972), Generalized Linear Mixed-effects Models (GLMMs, Jamil & ter Braak 2013), or more generally the Maximum entropy framework (Phillips *et al.* 2006; Elith *et al.* 2011). However, recently models that can be fitted to data from multiple species simultaneously, such as multivariate Generalized Linear Models (MGLMs, Wang *et al.* 2012) have increased in popularity. MGLMs can estimate parameters for the niche of all species simultaneously, which is more convenient than fitting a single model to each species separately as in more standard species distribution models (Warton *et al.* 2015b).

Similarly to the difference between GLMs and mixed-effects models, MGLMs do not account for residual variation unaccounted for by the predictors which can arise due to co-occurrence of species, and thus needs to be addressed to improve model fit. Generalized Linear Latent Variable Models (GLLVMs, Warton *et al.* 2015a; Ovaskainen *et al.* 2017; Niku *et al.* 2019) are Joint Species Distribution Models (JSDMs, Pollock *et al.* 2014; Clark *et al.* 2014) in that they account for species co-occurrence patterns by modelling patterns of residual variation. Unlike ordinary JSDMs, the residual covariance matrix is modelled in reduced-rank form, so that it includes fewer parameters, and so that the models are feasible to fit even for a large number of species, though there are alternative approaches without latent variables (Pichler & Hartig 2021).

Since GLLVMs are fitted to all species in a dataset at the same time, the GLLVM framework naturally lends itself for the calculation of niche overlap. In this article we present a model-based measure of niche overlap, similar to that of Swanson *et al.* (2015) in interpretation: the probability of observing a species given the predicted niche of another. The proposed measure of niche overlap is calculated by first fitting a model to estimate parameters for species niches, and secondly by predicting with the same model. Although we demonstrate the method using GLLVMs here, it is suitable to apply using any regression method, such as using a series of univariate models. Calculating niche overlap by predicting from a (multivariate) regression provides various opportunities, including; the prediction of niche overlap for unobserved values of the environment, niche overlap for multiple species, and the calculation of confidence intervals for niche overlap. We demonstrate the proposed model-based niche overlap measure by fitting a GLLVM with constrained latent variables and quadratic response model to abundances of Foramnifera species in the Spermonde archipelago, Indonesia (Cleary & Renema 2007).

Methods

In this section we first present the proposed model-based measure of niche overlap, after which we present the model we here use to predict species niches, and lastly we describe the data used in the example.

A species niche

Per species, a GLLVM represents a hypersurface in $k = 1 \dots K$ dimensions, such as with e.g., measured predictor variables, latent variables, ordination axes, or similar. For a single species $j = 1 \dots p$, let the following quantity denote the multidimensional niche volume:

$$|j| = \int \mathrm{F}_j(\boldsymbol{\lambda}) p(\boldsymbol{\lambda}) d\boldsymbol{\lambda}. \tag{1}$$

where $\mathrm{F}_j(\cdot)$ is a generic function representing the multidimensional niche of species j in a regression, and $p(\cdot)$ the probability of observing a particular environment. The likelihood of observing a species naturally depends on the likelihood of observing suitable environmental conditions, hence we additionally weigh a species niche by the probability of the environment. This can be the proportions of various environmental conditions in a measured predictor variable, or alternatively, from a larger body of data such as a digital elevation model.

The quantity in equation (1) forms the normalising constant for the calculation of a species' probability of occurrence anywhere in its niche, so that any species' probability to occur inside its own niche is one, and the probability for any species to occur outside of its niche is zero. For quadratic curves as in van der Veen *et al.* (2021b), and in a single dimension, this quantity implies an area, whereas in two dimensions it implies a volume, so that in three or more dimensions the ecological niche is implicitly represented as a hypervolume. Similarly, the number of dimensions K can be considered the number of niche axes (Hutchinson & MacArthur 1959), representing e.g., soil properties of a grassland, or the amount of precipitation in a given time period. For measured predictor variables in a regression, the solution to equation (1) can be straightforwardly calculated by averaging over the predictions of a regression.

Niche overlap

Similar to the probability of observing a single species, the probability of observing two (or more) species j, m together, given that species j has been observed, becomes smaller when species j is predicted occur infrequently (and naturally even more so if species m is predicted to occur infrequently). As such, the model-

based measure of niche overlap presented here is small when: 1) species j is predicted to occur infrequently, 2) species m is predicted to occur infrequently, or 3) the environment where species j and m are predicted to both occur is rarely observed. A species can be predicted to occur infrequently for a variety of reasons, such as a low mean abundance, or due to generally negative estimated responses to predictors in the regression. Ecologically, the reasons for a species to occur infrequently are many, but can include e.g., a narrow niche (few suitable environments, i.e., it is a habitat specialist), or a low maximum of the niche (the species does not occur in large quantities, i.e. it is rare).

We define the proportion of the niche of species j that overlaps with the niche of species m as:

$$p(O_{j,m}) = \frac{\int \min(\mathrm{F}_j(\boldsymbol{\lambda}), \mathrm{F}_m(\boldsymbol{\lambda}))d\boldsymbol{\lambda}}{|j|} \tag{2}$$

where $\min(\cdot)$ indicates the minimum space of the response curves $\mathrm{F}_j(\cdot)$ and $\mathrm{F}_m(\cdot)$, which is integrated over, as visualized in Figure 1, and visualized from a one-dimensional perspective in Figure 2.
This measure can be thought of as the probability of observing both species, given the predicted niche of one, or as the probability of observing the environment suitable for both species, given that one of the species has been observed. This measure is similar to species associations represented by residual covariances of a JSDM. However, most commonly, species associations from a JSDM are calculated without accounting for other effects in a model, such as environmental predictors or species constants that account for the frequency at which a species is predicted to occur (e.g. the intercept, or alternatively excluding species mean abundances) (Pollock *et al.* 2014). Additionally, species associations from JSDMs can only be calculated for two species, whereas the measure presented here can be extended to include many more species, up to p, as we demonstrate in the Foraminifera example below.

Predicting niche overlap

Taking a model-based approach to calculating niche overlap has various benefits. In general, any regression method can be used to make predictions using the parameter estimates. This can be for an environment that has been observed, or alternatively, new measurements of the environment can be provided to predict niche overlap in places that have not yet been observed. For example, we might record species on an elevation gradient from 100 meters above sea level to 400 meters above sea level, and at 10 - 15 degrees Celsius. Then, with species responses to elevation and temperature estimated in the model, we can additionally predict niche overlap at 1000 meters above sea level and at 16 degrees Celsius.

Similarly, by keeping specific predictors in the model constant, e.g., by fixing a predictor at a certain level or value, a conditional niche overlap measure is retrieved. Imagine that besides elevation a predictor

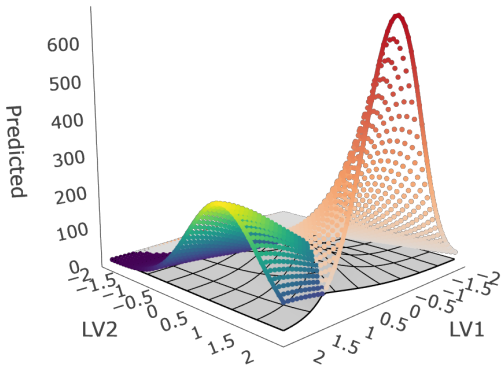


Figure 1: Three dimensional figure of two species quadratic niches, predicted using the Foraminifera dataset below. The predicted niche overlap is conditional, for when Foraminifera were not exposed to oceanic current, and for a depth of zero. Dots form the outline of the niches, with colors indicating change in predicted abundances. The grey surface outlines the volume that represents the overlap between the two species, corresponding with the methodology explained above.

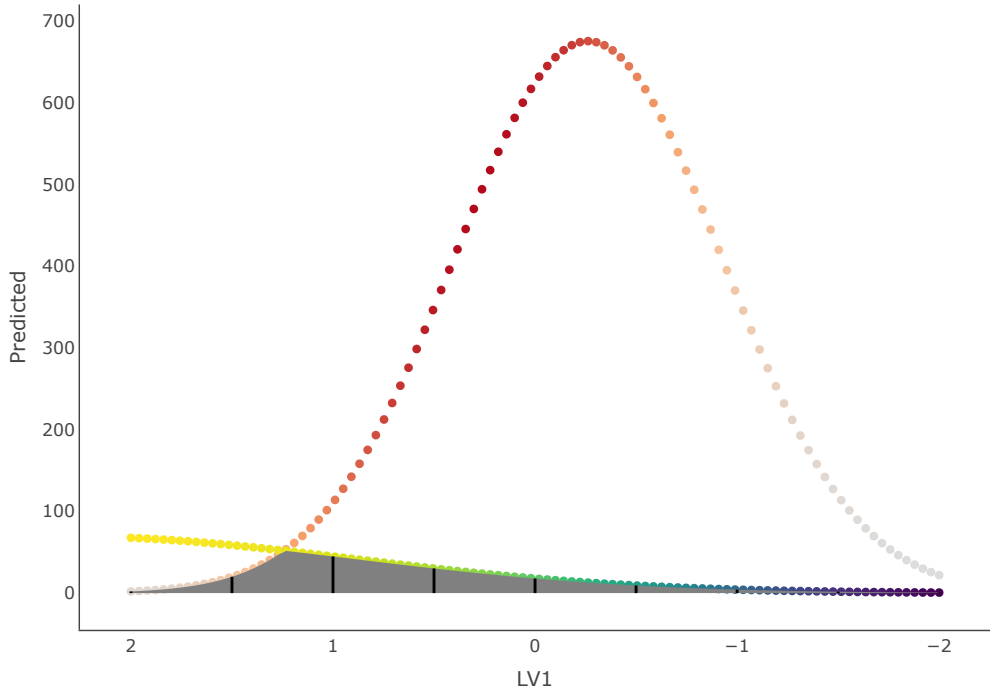


Figure 2: One-dimensional perspective of Figure 1, for latent variable 1. Latent variable two has been fixed at 2, so that the view provided is that at the edge of Figure 1. The niche overlap surface is now seen as an area (the space under the gray surface in Figure 1). The dots for each species indicate the predicted surface for a single species' niche, the same as in figure 1.

for management type is included in the model, with the levels "frequent" and "infrequent". For a quadratic response curve, each category represents a separate maximum of the niche, so that niche overlap will be different for both categories. It is possible to average over both effects as a means to retrieve an overall measure of niche overlap, or two conditional measures of niche overlap can be calculated: one for each category of the predictor.

The integration in equation (1) and equation (2) in most cases does not have an analytical solution. Thus, we use numerical integration to approximate the solution. For computational reasons, here we take a simulation-based approach. Note, that simulating from predictors is equivalent to predicting niche overlap for unobserved combinations of the environment. An alternative could be to calculate niche overlap using the observed site conditions, to represent "observed" niche overlap, though such a measure would be highly susceptible to variation in the sampling of sites. For categorical predictors we sample the observed categories with a probability equal to their proportion in the data. We simulate continuous predictors from a uniform distribution, with the minimum and maximum of each simulation equal to those in the data, i.e., by default we integrate over the range of the observed environment, rather than all possible environments. Alternatively, in the case of predictors for which data is available on a larger scale, as is the case for e.g., elevation from a Digital Elevation Model, or from world-wide observations for temperature, one could instead simulate from a larger body of data, and calculate the predicted niche overlap accordingly.

We represent the probability of the environment for continuous predictors $\mathrm{p}(\lambda_k)$ using kernel density estimates (using the `density` function in the `stats` R-package), independently for all predictors. Alternatively, a multivariate kernel density could be considered, though we consider that an avenue for future research. Then, we use importance sampling to more accurately perform the integration, and to evaluate the integrand more often in places of the environment that are more frequently sampled. We determine the probability of unobserved values for predictors by linear interpolation (using the `approxfun` function in the `stats` R-package), so that we can re-weight each of the $r = 1 \dots R$ realizations from the uniform distributions by their (approximate) probabilities. Lastly, we average over the result:

$$p(O_{j,m}) \approx \frac{1}{R} \sum_{r=1}^{R} \frac{\min(\mathrm{F}_j(\boldsymbol{\lambda}_r), \mathrm{F}_m(\boldsymbol{\lambda}_r))}{|j|} \prod_{k=1}^{K} \frac{p(\lambda_{kr})}{\mathrm{Unif}(\lambda_{kr})}. \tag{3}$$

When random effects are additionally included in the model (as is the case for GLLVMs), one needs to separately consider how to treat these. For example, it is possible to predict using the conditional distribution of the random effects (e.g., as in Hui *et al.* 2017), or one simulate from the marginal distribution of the random effect. For unobserved combinations of predictors, it is not possible to predict using the conditional distribution of the random effects, so that it is necessary to simulate from the marginal distribution of the

random effect (and average over the predictions), which is usually assumed to be normally distributed. The latter approach combines especially well with the model-based measure of niche overlap proposed here, as it allows for straightforward implementation of the integration in equation (2).

We demonstrate predicting (conditional) niche overlap for multiple species in the Foraminifera example below.

Confidence intervals for niche overlap by bootstrap

Since we are predicting niche overlap, presenting a statistical uncertainty of that prediction is vital for a thorough representation of the likelihood of niche overlap for two or more species. This can be represented using a confidence interval for the prediction, though this is difficult to calculate. As such, we use the estimated variance-covariance matrix of parameter estimates to simulate from the asymptotic (multivariate normal) distribution of parameter estimates. In short, we simulate new parameter estimates, predict niche overlap, and repeat S times to retrieve a distribution for the niche overlap between all species. Then, we use the 2.5 and 97.5% percentiles from that distribution to represent a confidence interval for the proposed model-based measure of niche overlap.

Case study

To demonstrate the proposed model-based measure of niche overlap we use a dataset of counts from $p = 22$ species of Foraminifera at $n = 31$ locations (Becking *et al.* 2006; Cleary & Renema 2007). The dataset is publicly available as part of the CESTES database (Jeliazkov *et al.* 2020). Large benthic Foraminifera were collected at the Spermonde Archipelago in Indonesia, by scubadiving at various locations, during the period of July-October in 1997. Various environmental drivers were recorded, including the depth at which the sample was collected, the presence of sedimentary areas, coral formations, reef flats, distance to the nearest human settlement, visibility, exposure of the sample to oceanic swell, and the micro-substrate on which each sample was found. Cleary & Renema (2007) analysed the data using Canonical Correspondence Analysis (CCA, ter Braak 1986). Cleary & Renema (2007) concluded that depth, micro-substrate, and visibility were the most important drivers of Foraminifera distribution. Unfortunately, the micro-substrate measurements were not available in the CESTES database, so that we were unable to include that here. Though species names were not included in the database, we used the total number of counts from each species, reported in the original study, and species traits, to assign the species names (see Appendix S1 for a full list of species names and abbreviations). We excluded *Laevipeneroplis malayensis* and *Operculina complanata*, as those species had few observations.

The exposure variable included four different categories, representing how exposed a sample was to ocean currents, but we grouped this into two categories due to insufficient sample sizes: not exposed ($N = 16$) and exposed ($N = 15$).

The dataset contains large counts of Foraminifera species and thus ample information to fit a GLLVM with quadratic response model, so we here fitted a GLLVM with constrained latent variables and quadratic response model (but species-common tolerances, see van der Veen *et al.* (2021b)) as implemented in the `gllvm` R-package (Niku *et al.* 2020), to study co-occurrence patterns of Foraminifera species.

First, we calculate pairwise niche overlap with $R = 10^6$ realizations for the predictors and latent variables, and with $S = 1000$ simulations from the estimated covariance matrix of the parameter estimates to represent confidence intervals of the predicted niche overlap. Second, we calculate niche overlap for all combinations of overlapping species (e.g., starting at pairwise and ending at the niche overlap of all p species) with $R = 10^4$, which is lower than for the pairwise niche overlap for computational reasons, to study the mean trend of niche overlap. In total, there are $O = 2 \dots p \sum_{o=1}^{O} \binom{p}{o}$ combinations of overlapping species, so that per species there are $\sum_{o=1}^{O-1} \binom{p-1}{o}$ potential measures of niche overlap (for the dataset below, this is approximately four and two million respectively). Lastly, we present conditional niche overlap for some species as demonstration of the methodology, with $R = 10^6$ realizations for the predictors and latent variables, and with $S = 1000$ simulations from the estimated covariance matrix of the parameter estimates. The code that was used to calculate (pairwise) niche overlap is included in Appendix S2.

Results

The predicted pairwise niche overlap of Foraminifera species is presented in Figure 3. Each panel represents a Foraminifera species in the dataset, and each dot the predicted niche overlap by the model, for that species with another species in the dataset. Since each prediction is a function of the parameter estimates of two species, confidence intervals tend to be wide, as they represent uncertainty in the prediction due to both sets of parameter estimates. Regardless, it is clear that some species are predicted to overlap with few other species (e.g. *Operculina ammonoides*) and some species with many other species (e.g. *Amphistegina radiata*).

For two species, the asymmetric niche overlap measure is based on the niche overlap, but the normalization is different (as it is performed per species niche, so that the overlap can be for the niche of species 1 given the predicted niche of species two or for the niche of species two given the predicted niche of species one). Due to the large amount of information when calculating niche overlap for more than a few species, it is difficult to present in an informative manner. In Figure 4 we have chosen to visualize the average niche overlap for all species, and for an increasing number of overlapping species, though without confidence intervals

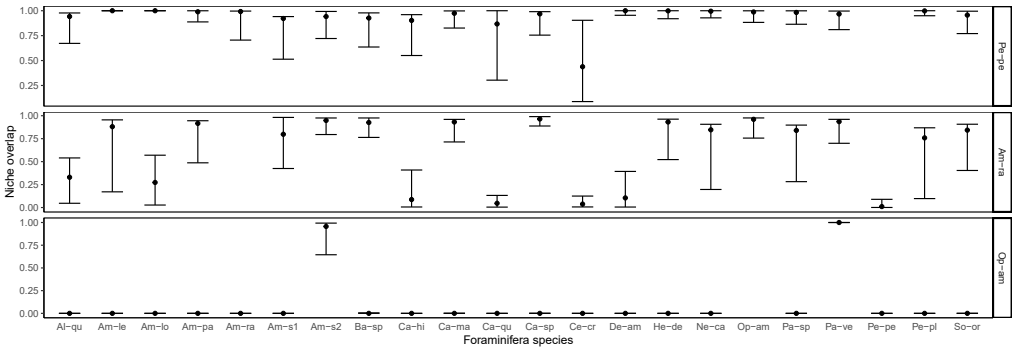


Figure 3: Pairwise niche overlap predicted with the GLLVM with constrained latent variables for three of the Foraminifera species. Dots represent the predicted niche overlap per species, and error bars represent 95% confidence intervals of the prediction, calculated by simulating parameter estimates from the covariance matrix of the fitted model. Each panel represents a single species, and each dot the overlap with other Foraminifera species as part of its niche. A plot including the pairwise niche overlap for all species is included in appendix S1.

(as these are computationally intensive to calculate using the simulation approach used here). Immediately when studying Figure 4, it becomes apparent that niche overlap decreases rapidly with the number of species involved in the calculation. Few species overlap with many other species, and as the figure shows, *Peneroplis pertusis* is the only species that is likely to be observed together with all other species.

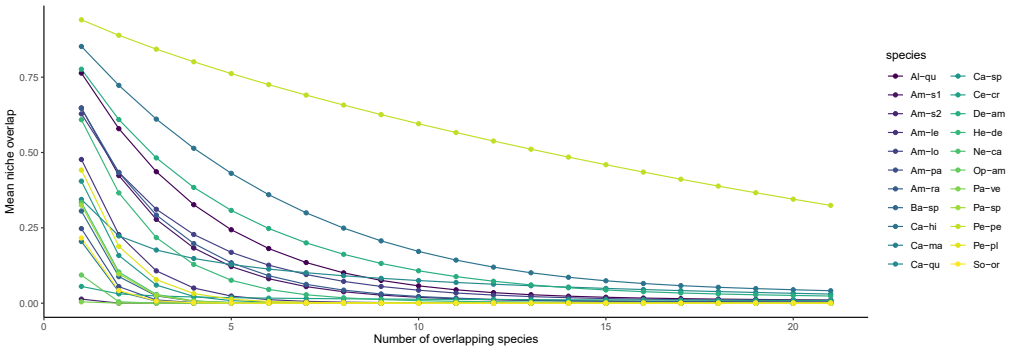


Figure 4: Average niche overlap between all combinations of species. Niche overlap decreases when calculated for an increasing number of overlapping species. Species that overlap with many other species will have a large average niche overlap measure.

The niche overlap measures in Figure 3 and Figure 4 represent niche overlap for the whole environment. Instead, we could have chosen to produce figures presenting niche overlap for e.g., environments at 30 meters depth, or for environments with a high degree of visibility. To demonstrate, the niche overlap of all 22 Foraminifera species for environments exposed to oceanic current, and as part of the niche of *Peneroplis*

pertusis is 0.41 (95% CI: 0.09, 0.66) in comparison to the 0.29 (0.05, 0.52) for environments not exposed to oceanic currents. Thus, for this example it is not possible to conclude that the predicted niche overlap differed between exposed and not exposed environments.

Discussion

In this article, we present a model-based measure of niche overlap, which can aid researchers in better understanding species co-occurrence patterns in ecological communities, for example by allowing for the identification of indicator species (Warton *et al.* 2015a; Hui 2016). For a community of Foraminifera species, we demonstrated the potential applications of model-based niche overlap by fitting a recently developed method for model-based ordination with constrained latent variables (van der Veen *et al.* 2021a). However, it is possible to predict niche overlap using any method for regression e.g., with stacked species distributions models potentially using the Maximum Entropy framework (Phillips *et al.* 2006), from JSDMs (Pollock *et al.* 2014), or more generally from multispecies models instead.

In this article, species' responses were presented with a relatively simple quadratic function. But, many ecologists make use of more complex functions to describe species niches, for example by fitting Generalized Additive Models (GAMs, Wood 2017) instead. GAMs allow researchers to specify flexible response curves in order to model species responses to the environment in more flexibly. Since species co-occurrence patterns are inherently multidimensional, specifying species responses to the environment separately for each dimension can make it difficult to explore species niches, increasing the potential to miss out on important patterns for (lack of) niche overlap. In contrast, the measure of niche overlap presented here always includes all drivers of co-occurrence patterns in a model. Additionally, this has the benefit of an explicit connection with classical niche overlap theory, which many ecologists are taught during their studies. Measures of niche overlap have been used in ecology for decades (Pianka 1973; Hurlbert 1978), so that presenting species co-occurrence as niche overlap has the potential to provide ecologists with a sense of familiarity.

Since all information in the model is condensed into a single, straightforward to interpret measure, the approach presented here facilitates a more unified inference on co-occurrence patterns in ecological communities. In contrast, models that additionally include random effects, such as JSDMs or mixed-models in general, complicate drawing complete inference on species co-occurrence patterns even further. For predictors, a shared environmental response is identified by similar parameter estimates for predictor slopes of species. However, in JSDMs additional unmeasured patterns of species co-occurrence are represented by the residual correlation matrix (Warton *et al.* 2015a). There has been a growing concern about the interpretation of residual correlations as evidence of species interactions (Blanchet *et al.* 2020). Inferring

species co-occurrence through a measure of niche overlap might turn the tide in that argument, since it can trigger ecologists to more often consider shared environmental responses as a reason for species co-occurrence instead.

Most methods for the calculation of niche overlap focus on pairwise combinations of species (e.g. Geange *et al.* 2011; Blonder *et al.* 2014), while pairwise niche overlap can be affected by all species in the community As such, the choice of focussing on pairwise niche overlap does not reflect the degree of complexity expected in a real ecological community. Instead, in this article, we predicted niche overlap for all potential combinations of overlapping species, starting at pairwise and finally calculating niche overlap for all species in the community, and separately for different environments. For example, for the Foraminifera dataset the results here indicate that when an individual of *Peneroplis pertusis* has been observed, it is likely to also observe individuals of all other Foraminifera species in the dataset, whereas we could not have drawn that conclusion had we only studied pairwise niche overlap. This pattern did not seem to differ between environments that were exposed to oceanic currents with various degree. The ability to examine niche overlap between more than two species, and to explore how niche overlap changes along environmental gradients, or how it is influenced by environmental change, are interesting features of the method presented in this article with potential for community ecology.

With the possibility of calculating niche overlap for all combinations of all species in a community comes the need to make decisions about what results to present. Here, we chose to present pairwise niche overlap for historical reasons, and the average predicted niche overlap for an increasing number of species to identify any general trends in niche overlap. Due to the large number of possible results, somewhat arbitrary choices need to be made for the presentation of niche overlap in ecological studies on a case-by-case basis. Further research could attempt to establish a best practice for the presentation of niche overlap, but also on improving the software implementation and the methodology presented here. Though it is possible to calculate niche overlap for all combinations of species using sufficient computing resources, numerical integration by simulation is computationally intensive. In practice, the calculation of model-based niche overlap presented here is restricted by the number of predictors, random effects, species, and simulations in the analysis. Thus, more efficient methods for computation should be considered in future studies, such as Laplace's method.

Acknowledgements

B.V. was supported by a scholarship from the Research Council of Norway (grant number 272408/F40). F.K.C.H was supported by an Australia Research Council Discovery Fellowship (grant number

DE200100435).

References

Austin, M.P. (1987). Models for the analysis of species' response to environmental gradients. *Vegetatio*, **69**, 35–45.

Austin, M.P., Nicholls, A.O. & Margules, C.R. (1990). Measurement of the Realized Qualitative Niche: Environmental Niches of Five Eucalyptus Species. *Ecological Monographs*, **60**, 161–177.

Becking, L.E., Cleary, D.F.R., Voogd, N.J. de, Renema, W., Beer, M. de, Soest, R.W.M. van & Hoeksema, B.W. (2006). Beta diversity of tropical marine benthic assemblages in the Spermonde Archipelago, Indonesia. *Marine Ecology*, **27**, 76–88.

Blanchet, F.G., Cazelles, K. & Gravel, D. (2020). Co-occurrence is not evidence of ecological interactions. *Ecology Letters*, **23**, 1050–1063.

Blonder, B., Lamanna, C., Violle, C. & Enquist, B.J. (2014). The n-dimensional hypervolume. *Global Ecology and Biogeography*, **23**, 595–609.

Cattin, M.-F., Bersier, L.-F., Banašek-Richter, C., Baltensperger, R. & Gabriel, J.-P. (2004). Phylogenetic constraints and adaptation explain food-web structure. *Nature*, **427**, 835–839.

Chase, J.M. & Leibold, M.A. (2003). *Ecological Niches: Linking Classical and Contemporary Approaches*. University of Chicago Press.

Chesson, P. (2000). Mechanisms of Maintenance of Species Diversity. *Annual Review of Ecology and Systematics*, **31**, 343–366.

Clark, J.S., Bell, D.M., Kwit, M.C. & Zhu, K. (2014). Competition-interaction landscapes for the joint response of forests to climate change. *Glob Chang Biol*, **20**, 1979–1991.

Cleary, D.F.R. & Renema, W. (2007). Relating species traits of foraminifera to environmental variables in the Spermonde Archipelago, Indonesia. *Marine Ecology Progress Series*, **334**, 73–82.

Cohen, J.E. (1978). *Food Webs and Niche Space. (MPB-11), Volume 11*. Princeton University Press.

Cohen, J.E. (1977). Food webs and the dimensionality of trophic niche space. *PNAS*, **74**, 4533–4536.

Darwin, C. (1859). *The origin of species; and, the descent of man*. Modern library.

Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E. & Yates, C.J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, **17**, 43–57.

Elton, C.S. (1927). *Animal ecology*. Macmillan Co., New York,.

Gause, G. (1934). The struggle for existence. *Baltimore, Maryland*.

Geange, S.W., Pledger, S., Burns, K.C. & Shima, J.S. (2011). A unified analysis of niche overlap incorporating data of different types. *Methods in Ecology and Evolution*, **2**, 175–184.

Godsoe, W. & Harmon, L.J. (2012). How do species interactions affect species distribution models? *Ecography*, **35**, 811–820.

Grinnell, J. (1924). Geography and Evolution. *Ecology*, **5**, 225–229.

Guisan, A. & Zimmermann, N.E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147–186.

Hardin, G. (1960). The Competitive Exclusion Principle. *Science*, **131**, 1292–1297.

Holt, R.D. (1987). On the Relation between Niche Overlap and Competition: The Effect of Incommensurable Niche Dimensions. *Oikos*, **48**, 110–114.

Hui, F.K.C. (2016). Boral – Bayesian Ordination and Regression Analysis of Multivariate Abundance Data in r. *Methods in Ecology and Evolution*, **7**, 744–750.

Hui, F.K.C., Warton, D.I., Ormerod, J.T., Haapaniemi, V. & Taskinen, S. (2017). Variational Approximations for Generalized Linear Latent Variable Models. *Journal of Computational and Graphical Statistics*, **26**, 35–43.

Hurlbert, S.H. (1978). The Measurement of Niche Overlap and Some Relatives. *Ecology*, **59**, 67–77.

Hutchinson, G.E. (1957). Concluding Remarks. *Cold Spring Harb Symp Quant Biol*, **22**, 415–427.

Hutchinson, G.E. (1959). Homage to Santa Rosalia or why are there so many kinds of animals? *The American Naturalist*, **93**, 145–159.

Hutchinson, G.E. & MacArthur, R.H. (1959). A Theoretical Ecological Model of Size Distributions Among Species of Animals. *The American Naturalist*, **93**, 117–125.

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

Jansen, F. & Oksanen, J. (2013). How to model species responses along ecological gradients – Huisman–Olff–Fresco models revisited. *Journal of Vegetation Science*, **24**, 1108–1117.

Jeliazkov, A., Mijatovic, D., Chantepie, S., Andrew, N., Arlettaz, R., Barbaro, L., Barsoum, N., Bartonova, A., Belskaya, E., Bonada, N., Brind'Amour, A., Carvalho, R., Castro, H., Chmura, D., Choler, P., Chong-Seng, K., Cleary, D., Cormont, A., Cornwell, W., de Campos, R., de Voogd, N., Doledec, S., Drew, J., Dziock, F., Eallonardo, A., Edgar, M.J., Farneda, F., Hernandez, D.F., Frenette-Dussault, C., Fried, G., Gallardo, B., Gibb, H., Gonçalves-Souza, T., Higuti, J., Humbert, J.-Y., Krasnov, B.R., Saux, E.L., Lindo, Z., Lopez-Baucells, A., Lowe, E., Marteinsdottir, B., Martens, K., Meffert, P., Mellado-Díaz, A., Menz, M.H.M., Meyer, C.F.J., Miranda, J.R., Mouillot, D., Ossola, A., Pakeman, R., Pavoine, S., Pekin, B., Pino, J., Pocheville, A., Pomati, F., Poschlod, P., Prentice, H.C., Purschke, O., Raevel, V., Reitalu,

T., Renema, W., Ribera, I., Robinson, N., Robroek, B., Rocha, R., Shieh, S.-H., Spake, R., Staniaszek-Kik, M., Stanko, M., Tejerina-Garro, F.L., ter Braak, C., Urban, M.C., van Klink, R., Villéger, S., Wegman, R., Westgate, M.J., Wolff, J., Żarnowiec, J., Zolotarev, M. & Chase, J.M. (2020). A global database for metacommunity ecology, integrating species, traits, environment and space. *Scientific Data*, **7**, 6.

Letten, A.D., Ke, P.-J. & Fukami, T. (2017). Linking modern coexistence theory and contemporary niche theory. *Ecological Monographs*, **87**, 161–177.

MacArthur, R. & Levins, R. (1967). The limiting similarity, convergence, and divergence of coexisting species. *The American Naturalist*, **101**, 377–385.

May, R.M. & Arthur, R.H.M. (1972). Niche Overlap as a Function of Environmental Variability. *Proceedings of the National Academy of Sciences of the United States of America*, **69**, 1109–1113.

McGill, B.J., Enquist, B.J., Weiher, E. & Westoby, M. (2006). Rebuilding community ecology from functional traits. *Trends in Ecology & Evolution*, **21**, 178–185.

Nelder, J.A. & Wedderburn, R.W.M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society: Series A (General)*, **135**, 370–384.

Niku, J., Brooks, W., Herliansyah, R., Hui, F.K.C., Taskinen, S., Warton, D.I. & van der Veen, B. (2020). *Gllvm: Generalized linear latent variable models*.

Niku, J., Hui, F.K.C., Taskinen, S. & Warton, D.I. (2019). Gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in r. *Methods in Ecology and Evolution*, **10**, 2173–2182.

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

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

Pianka, E.R. (1974). Niche Overlap and Diffuse Competition. *PNAS*, **71**, 2141–2145.

Pianka, E.R. (1973). The Structure of Lizard Communities. *Annual Review of Ecology and Systematics*, **4**, 53–74.

Pichler, M. & Hartig, F. (2021). A new method for faster and more accurate inference of species associations from big community data. URL http://arxiv.org/abs/2003.05331 [accessed 14 July 2021]

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

Swanson, H.K., Lysy, M., Power, M., Stasko, A.D., Johnson, J.D. & Reist, J.D. (2015). A new probabilistic method for quantifying n-dimensional ecological niches and niche overlap. *Ecology*, **96**, 318–324.

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

van der Veen, B., Hui, F.K.C., Hovstad, K.A. & O'Hara, R.B. (2021a). Model-based ordination with constrained latent variables.

van der Veen, B., Hui, F.K.C., Hovstad, K.A., Solbu, E.B. & O'Hara, R.B. (2021b). Model-based ordination for species with unequal niche widths. *Methods in Ecology and Evolution*, **n/a**.

Vogel, J.T., Somers, M.J. & Venter, J.A. (2019). Niche overlap and dietary resource partitioning in an African large carnivore guild. *Journal of Zoology*, **309**, 212–223.

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

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

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

Wathne, J.A., Haug, T. & Lydersen, C. (2000). Prey preference and niche overlap of ringed seals Phoca hispida and harp seals P. Groenlandica in the Barents Sea. *Marine Ecology Progress Series*, **194**, 233–239.

Wood, S.N. (2017). *Generalized Additive Models: An Introduction with R, Second Edition*. CRC Press.

Model-based analysis of niche overlap with Generalized Linear Latent Variable Models

Bert van der Veen[123] Robert B. O'Hara[23] Francis K.C. Hui[4]
Knut A. Hovstad[35]

[1]Department of Landscape and Biodiversity, Norwegian Institute of Bioeconomy research, Trondheim, Norway
[2]Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway
[3]Centre of Biodiversity Dynamics, Norwegian University of Science and Technology, Trondheim, Norway
[4]Research School of Finance, Actuarial Studies and Statistics, The Australian National University, Canberra, Australia
[5]The Norwegian Biodiversity Information Centre, Trondheim, Norway

Appendix S1

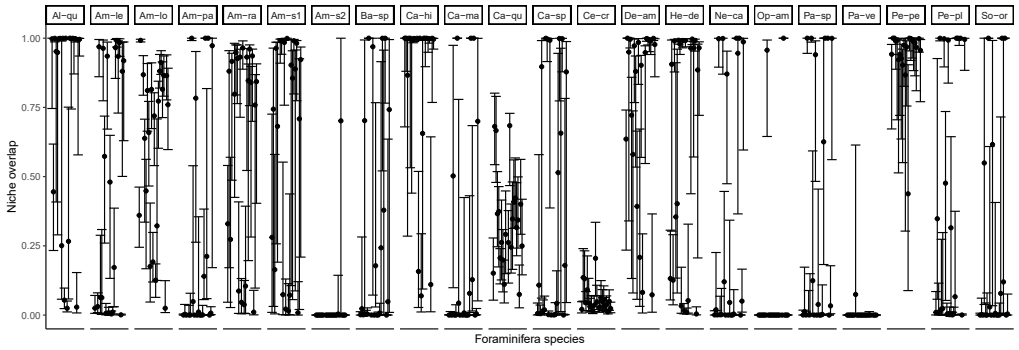
Table of species names

Species name	abbreviation
Calcarina spengleri	Al-qu
Calcarina mayori	Am-s1
Calcarina quoyii	Am-s2
Calcarina hispida	Am-le
Neorotalia calcar	Am-lo
Baculogypsinoides spinosus	Am-pa

(continued)

Species name	abbreviation
Amphistegina lessonii	Am-ra
Amphistegina lobifera	Ba-sp
Amphistegina papillosa	Ca-hi
Amphistegina radiata	Ca-ma
Operculina ammonoides	Ca-qu
Operculina complanata	Ca-sp
Heterostegina depressa	Ce-cr
Palaeonummulites venosus	De-am
Celanthus craticulatum	He-de
Peneroplis pertusus	
Peneroplis planatus	Ne-ca
Dendritina ambigua	Op-am
Alveolinella quoyii	
Laevipeneroplis malayensis	Pa-ve
Amphisorus sp 1	Pa-sp
Amphisorus sp 2	Pe-pe
Sorites orbiculus	Pe-pl
Parasorites sp 2	So-or

Figure S1: pairwise niche overlap of all species



Appendix S2: R-code used to calculate niche overlap

Function used to calculate pairwise niche overlap.

```
simOverlap <- function(object = 1000, nsim = 1000, newX = NULL, boot.CI = TRUE,
    bootSim = 1000, seed = NULL, ...) {
    set.seed(seed)
    # ld <- try(library('cubature'),silent=T) if(inherits(ld,'try-error')){
    # stop('Cubature R-package not installed.') }
    if (!is.list(object$sd) & boot.CI) {
        stop("Cannot calculate niche overlap without standard errors.")
    }
    if (object$family == "ordinal") {
        stop("Not implemented for ordinal model.")
    }

    p <- ncol(object$y)
    n <- nrow(object$y)
    if (!is.null(newX)) {
        n <- nrow(newX)
    }

    # Extract used predictors
    is.cat <- function(x) ifelse(all(x %in% c(0, 1)), TRUE, FALSE)
    if (is.null(object$lv.formula)) {
        object$lv.formula <- ~1
    }
    if (!is.null(cbind(object$X.design, object$lv.X))) {
        X <- cbind(object$X, object$lv.X)

        X <- X[, unique(c(colnames(object$params$Xcoef), row.names(object$params$LvXcoef)))]
    } else {
        X <- NULL
```

```
}
if (boot.CI) {

    # Covariance matrix of parameters
    Sigma_pars <- try(solve(object$Hess$Hess.full[object$Hess$incl,
        object$Hess$incl]), silent = T)
    if (inherits(Sigma_pars, "try-error")) {
        Sigma_pars <- try(MASS::ginv(object$Hess$Hess.full[object$Hess$incl,
            object$Hess$incl], tol = 0), silent = T)
    }
    if (inherits(Sigma_pars, "try-errors")) {
        Sigma_pars <- object$Hess$cov.mat.mod
    }
    # more robust way of sampling from MVRNORM than using Sigma directly
    L <- suppressWarnings(try(chol(Sigma_pars), silent = T))
    if (inherits(L, "try-error")) {
        if (min(diag(Sigma_pars)) > -0.01) {
            # add small value to perturb the covariance matrix and try again
            L <- suppressWarnings(try(chol(Sigma_pars + abs(min(diag(Sigma_pars))) +
              1e-08), silent = T))
            if (inherits(L, "try-error")) {
              L <- suppressWarnings(try(chol(Sigma_pars, pivot = T),
                silent = T))
              if (inherits(L, "try-error")) {
                stop("Non-singular covariance matrix of parameters.")
              }
            }
        } else if (inherits(L, "try-error")) {
            L <- suppressWarnings(try(chol(Sigma_pars, pivot = T),
              silent = T))
        } else if (inherits(L, "try-error")) {
            stop("Non-singular covariance matrix of parameters.")
```

```
        }


    }


    # true parameters
    par <- object$TMBfn$par[object$Hess$incl]
}
# simulate newX
if (!is.null(X)) {
    newX <- matrix(0, ncol = ncol(X), nrow = nsim)
    for (i in 1:ncol(X)) {
        if (!is.cat(X[, i])) {
            # sample from uniform is continuous
            newX[, i] <- runif(nsim, min = min(X[, i]), max = max(X[,
              i]))
        } else {
            # re-sample factors or characters with prop as prob
            newX[, i] <- sample(unique(X[, i]), prob = table(X[, i])/nrow(X),
              size = nsim, replace = T)
        }


    }
    colnames(newX) <- colnames(X)
    # calculate probabilities of environment importance sampling for
    # continuous predictors
    probs <- NULL
    for (i in which(!apply(X, 2, is.cat))) {
        densfun <- approxfun(density(X[, i], from = min(X[, i]), to = max(X[,
            i])), yleft = 0, yright = 0)
        probs <- cbind(probs, densfun(newX[, i])/dunif(newX[, i], min(X[,
            i]), max(X[, i])))
    }
```

```
} else {
    probs <- matrix(1, ncol = 1)
    newX <- NULL
}
# Simulate from MVRNORM for LVs
if (!is.null(object$lvs)) {
    newLV <- matrix(rnorm(nsim * ncol(object$lvs)), ncol = ncol(object$lvs))
    # If constrained ord, ensure independence of the LV
    if (object$num.lv.c > 0) {
        newLV[, 1:object$num.lv.c] <- residuals.lm(lm(newLV ~ newX))
    }
} else {
    newLV <- NULL
}

# predict
preds <- predict.gllvm(object, newX = newX, newLV = newLV, type = "response")
# Norm constant per niche
probs <- rowProds(probs)
speciesArea <- apply(preds * probs, 2, mean)

# Calculate overlap
overlap <- matrix(0, p, p)
for (j2 in 1:(p - 1)) {
    for (j in (j2 + 1):p) {
        overlap[j, j2] = mean(rowMins(preds[, c(j, j2)] * probs))
        overlap[j2, j] = overlap[j, j2]/speciesArea[j2]
        overlap[j, j2] = overlap[j, j2]/speciesArea[j]
    }
}
# bootstrap a CI
if (boot.CI) {
```

```
sim_overlap <- array(0, dim = c(bootSim, p, p))
progress_bar = txtProgressBar(min = 0, max = bootSim, style = 3,
    char = "-")
setTxtProgressBar(progress_bar, value = 0)

for (i in 1:bootSim) {
    # start simulation

    # First simulate from a std. normal distribution
    pars <- rnorm(length(par))
    # then transform with cholesky and means
    pars <- pars %*% t(L) + par

    # Now we start assigning parameters to the model again
    names(pars) <- names(object$TMBfn$par)[object$Hess$incl]

    beta <- pars[names(pars) == "b"]

    # intercepts
    object$params$beta0 <- beta[1:p]
    beta <- beta[-c(1:p)]

    # Slopes for predictors if present
    if (length(beta) > 0) {
        object$params$Xcoef <- matrix(beta, nrow = p)
    }

    # Slopes for reduced Rank
    if (object$num.RR > 0 | object$num.lv.c > 0) {
        b_lv <- pars[names(pars) == "b_lv"]
        object$params$LvXcoef <- matrix(b_lv, ncol = object$num.RR +
          object$num.lv.c)
```

```
}

# Slopes for LV if present
if (object$num.lv > 0 | object$num.lv.c > 0 | object$num.RR >
    0) {
    lambda <- matrix(0, nrow = p, ncol = object$num.lv + object$num.lv.c +
      object$num.RR)
    diag(lambda) <- 1
    lambda[lower.tri(lambda, diag = F)] <- pars[names(pars) ==
      "lambda"]
}

# Quadratic coefs for LVs
if (object$quadratic != F) {
    lambda2 <- pars[names(pars) == "lambda2"]
    lambda2 <- matrix(lambda2, nrow = p, ncol = object$num.lv +
      object$num.lv.c + object$num.RR, byrow = T)
    lambda <- cbind(lambda, lambda2)
}

# Assign the lambdas
if ((object$num.lv + object$num.lv.c + object$num.RR) > 0) {
    object$params$theta <- lambda
}

# Sigma for LVs
if ((object$num.lv + object$num.lv.c) > 0) {
    object$params$sigma.lv <- abs(pars[names(pars) == "sigmaLV"])
}

# Nuisance parameters
if (object$family == "ZIP") {
```

```
            lg_phi <- pars[names(pars) == "lg_phi"]
            object$params$inv.phi <- exp(lg_phi)
            object$params$phi <- 1/object$params$inv.phi
        }

        # Predict
        preds <- predict.gllvm(object, newX = newX, newLV = newLV,
            type = "response")
        # Norm constants
        speciesArea <- apply(preds * probs, 2, mean)

        # Calculate overlap
        for (j2 in 1:(p - 1)) {
            for (j in (j2 + 1):p) {
              sim_overlap[i, j, j2] = mean(rowMins(preds[, c(j, j2)] *
                probs))
              sim_overlap[i, j2, j] = sim_overlap[i, j, j2]/speciesArea[j2]
              sim_overlap[i, j, j2] = sim_overlap[i, j, j2]/speciesArea[j]
            }
        }
        setTxtProgressBar(progress_bar, value = i)

    }  #end simulation for CI

    # Assign everything to a dataframe
    specnam <- colnames(object$y)
    ovlp <- NULL
    for (j in 1:p) {
        for (j2 in (1:p)[-j]) {
            CI <- quantile(sim_overlap[, j, j2], c(0.025, 0.975), na.rm = T)  #overlap was median
            ovlp <- rbind(ovlp, cbind(overlap = overlap[j, j2], lower = CI[1],
              upper = CI[2], species = specnam[j], species2 = specnam[j2]))
```

```
            }
        }
        ovlp <- data.frame(ovlp)
        ovlp$overlap <- as.numeric(ovlp$overlap)
        ovlp$lower <- as.numeric(ovlp$lower)
        ovlp$upper <- as.numeric(ovlp$upper)
        close(progress_bar)
        # }
    } else {
        ovlp <- overlap
    }

    return(ovlp)
}
```

21 Function used to calculate p-wise nicheoverlap

```
simOverlapAny <- function(object = 1000, species.idx = NULL, nsim = 1000,
    newX = NULL, boot.CI = TRUE, bootSim = 1000, seed = NULL, ...) {
    set.seed(seed)
    if (is.null(species.idx)) {
        stop("Species.idx needs to be provided.")
    }
    # ld <- try(library('cubature'),silent=T) if(inherits(ld,'try-error')){
    # stop('Cubature R-package not installed.') }
    if (!is.list(object$sd) & boot.CI) {
        stop("Cannot calculate niche overlap without standard errors.")
    }
    if (object$family == "ordinal") {
        stop("Not implemented for ordinal model.")
    }
    if (!is.matrix(species.idx)) {
        species.idx <- matrix(species.idx, nrow = 1)
```

```
}
p <- nrow(species.idx)

n <- nrow(object$y)
if (!is.null(newX)) {
    n <- nrow(newX)
}

# Extract used predictors
is.cat <- function(x) ifelse(all(x %in% c(0, 1)), TRUE, FALSE)
if (is.null(object$lv.formula)) {
    object$lv.formula <- ~1
}

if (!is.null(cbind(object$X.design, object$lv.X))) {
    if (is.null(newX)) {
        X <- cbind(object$X, object$lv.X)

        X <- X[, unique(c(colnames(object$params$Xcoef), row.names(object$params$LvXcoef))),
            drop = F]
    } else {
        X <- newX

        X <- X[, unique(c(colnames(object$params$Xcoef), row.names(object$params$LvXcoef))),
            drop = F]
    }

} else {
    X <- NULL
}
if (boot.CI) {
    # Covariance matrix of parameters
```

```
Sigma_pars <- try(solve(object$Hess$Hess.full[object$Hess$incl,
    object$Hess$incl]), silent = T)
if (inherits(Sigma_pars, "try-error")) {
    Sigma_pars <- try(MASS::ginv(object$Hess$Hess.full[object$Hess$incl,
        object$Hess$incl], tol = 0), silent = T)
}
if (inherits(Sigma_pars, "try-errors")) {
    Sigma_pars <- object$Hess$cov.mat.mod
}
# more robust way of sampling from MVRNORM than using Sigma directly
L <- suppressWarnings(try(chol(Sigma_pars), silent = T))
if (inherits(L, "try-error")) {
    if (min(diag(Sigma_pars)) > -0.01) {
        # add small value to perturb the covariance matrix and try again
        L <- suppressWarnings(try(chol(Sigma_pars + abs(min(diag(Sigma_pars))) +
          1e-08), silent = T))
        if (inherits(L, "try-error")) {
          L <- suppressWarnings(try(chol(Sigma_pars, pivot = T),
            silent = T))
          if (inherits(L, "try-error")) {
            stop("Non-singular covariance matrix of parameters.")
          }
        }
    } else if (inherits(L, "try-error")) {
        L <- suppressWarnings(try(chol(Sigma_pars, pivot = T),
          silent = T))
    } else if (inherits(L, "try-error")) {
        stop("Non-singular covariance matrix of parameters.")
    }

}
```

```
    # true parameters
    par <- object$TMBfn$par[object$Hess$incl]
}
# simulate newX
if (!is.null(X)) {
    newX <- matrix(0, ncol = ncol(X), nrow = nsim)
    for (i in 1:ncol(X)) {
        if (!is.cat(X[, i])) {
            # sample from uniform is continuous
            newX[, i] <- runif(nsim, min = min(X[, i]), max = max(X[,
              i]))
        } else {
            # re-sample factors or characters with prop as prob
            newX[, i] <- sample(unique(X[, i]), prob = table(X[, i])/nrow(X),
              size = nsim, replace = T)
        }

    }
    colnames(newX) <- colnames(X)
    # calculate probabilities of environment importance sampling for
    # continuous predictors
    probs <- NULL
    for (i in which(!apply(X, 2, is.cat))) {
        densfun <- approxfun(density(X[, i], from = min(X[, i]), to = max(X[,
            i])), yleft = 0, yright = 0)
        probs <- cbind(probs, densfun(newX[, i])/dunif(newX[, i], min(X[,
            i]), max(X[, i])))
    }
} else {
    probs <- matrix(1, ncol = 1)
    newX <- NULL
}
```

```
# Simulate from MVRNORM for LVs
if (!is.null(object$lvs)) {
    newLV <- matrix(rnorm(nsim * ncol(object$lvs)), ncol = ncol(object$lvs))
    # If constrained ord, ensure independence of the LV
    if (object$num.lv.c > 0) {
        newLV[, 1:object$num.lv.c] <- residuals.lm(lm(newLV ~ newX))
    }
} else {
    newLV <- NULL
}

# predict could also do hacky predict trick here to speed up..
preds <- predict.gllvm(object, newX = newX, newLV = newLV, type = "response")
# Norm constant per niche
probs <- rowProds(probs)
speciesArea <- apply(preds * probs, 2, mean)

# Calculate overlap
overlap <- rep(0, p)
progress_bar = txtProgressBar(min = 0, max = p, style = 3, char = "-")
setTxtProgressBar(progress_bar, value = 0)

for (j in 1:p) {
    overlap[j] = mean(rowMins(preds[, species.idx[j, !is.na(species.idx[j,
        ])]] * probs))
    setTxtProgressBar(progress_bar, value = j)

    # /speciesArea[species.idx[j,!is.na(species.idx[j,])][1]]
}

# bootstrap a CI
if (boot.CI) {
```

```
speciesAreaold <- speciesArea[species.idx[, 1]]
sim_overlap <- matrix(0, nrow = bootSim, ncol = p)
progress_bar = txtProgressBar(min = 0, max = bootSim, style = 3,
    char = "-")
setTxtProgressBar(progress_bar, value = 0)

for (i in 1:bootSim) {
    # start simulation
    set.seed(i + seed)
    # First simulate from a std. normal distribution
    pars <- rnorm(length(par))
    # then transform with cholesky and means
    pars <- pars %*% t(L) + par

    # Now we start assigning parameters to the model again
    names(pars) <- names(object$TMBfn$par)[object$Hess$incl]

    # intercepts
    beta <- pars[names(pars) == "b"]
    object$params$beta0 <- beta[1:ncol(object$y)]

    # Slopes for predictors if present
    if (length(beta) > 0) {
        object$params$Xcoef <- matrix(beta, nrow = ncol(object$y))
    }

    # Slopes for reduced Rank
    if (object$num.RR > 0 | object$num.lv.c > 0) {
        b_lv <- pars[names(pars) == "b_lv"]
        object$params$LvXcoef <- matrix(b_lv, ncol = object$num.RR +
          object$num.lv.c)
    }
```

```
# Slopes for LV if present
if (object$num.lv > 0 | object$num.lv.c > 0 | object$num.RR >
    0) {
    lambda <- matrix(0, nrow = ncol(object$y), ncol = object$num.lv +
      object$num.lv.c + object$num.RR)
    diag(lambda) <- 1
    lambda[lower.tri(lambda, diag = F)] <- pars[names(pars) ==
      "lambda"]
}

# Quadratic coefs for LVs
if (object$quadratic != F) {
    lambda2 <- pars[names(pars) == "lambda2"]
    lambda2 <- matrix(lambda2, nrow = ncol(object$y), ncol = object$num.lv +
      object$num.lv.c + object$num.RR, byrow = T)
    lambda <- cbind(lambda, lambda2)
}

# Assign the lambdas
if ((object$num.lv + object$num.lv.c + object$num.RR) > 0) {
    object$params$theta <- lambda
}

# Sigma for LVs
if ((object$num.lv + object$num.lv.c) > 0) {
    object$params$sigma.lv <- abs(pars[names(pars) == "sigmaLV"])
}

# Nuisance parameters
if (object$family == "ZIP") {
    lg_phi <- pars[names(pars) == "lg_phi"]
    object$params$inv.phi <- exp(lg_phi)
```

```
                object$params$phi <- 1/object$params$inv.phi
            }

            # Predict
            preds <- predict.gllvm(object, newX = newX, newLV = newLV,
                type = "response")

            # Calculate overlap
            for (j in 1:p) {

                # Norm constants
                speciesArea <- mean(preds[, species.idx[j, 1]] * probs)
                sim_overlap[i, j] = mean(rowMins(preds[, species.idx[j,
                  !is.na(species.idx[j, ])]] * probs))/speciesArea
            }

            setTxtProgressBar(progress_bar, value = i)

        }  #end simulation for CI
    }

    if (boot.CI) {
        return(list(overlap = overlap/speciesAreaold, sim_overlap = sim_overlap))
    } else {
        return(list(overlap = overlap, speciesArea = speciesArea))
    }

}
```

2

GLLVMs in the real world

2.1 Modelling temperature-driven changes in species associations across freshwater communities

Received: 30 October 2021 | Accepted: 4 July 2021

DOI: 10.1111/gcb.15888

PRIMARY RESEARCH ARTICLE

Global Change Biology WILEY

Modelling temperature-driven changes in species associations across freshwater communities

Sam Wenaas Perrin[1] | Bert van der Veen[2,3] | Nick Golding[4,5,6] | Anders Gravbrøt Finstad[1]

[1]Centre of Biodiversity Dynamics, Department of Natural History, Norwegian University of Science and Technology, Trondheim, Norway

[2]Department of Landscape and Biodiversity, Norwegian Institute of Bioeconomy Research, Trondheim, Norway

[3]Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

[4]Telethon Kids Institute, Perth Children's Hospital, Nedlands, Western Australia, Australia

[5]Curtin University, Bentley, Western Australia, Australia

[6]Department of BioSciences, University of Melbourne, Parkville, Victoria, Australia

Correspondence
Sam W. Perrin, Centre of Biodiversity Dynamics, Department of Natural History, Norwegian University of Science and Technology, N-7491 Trondheim, Norway.
Email: sam.perrin@ntnu.no

Funding information
Norges Forskningsråd, Grant/Award Number: 243910 and 266574; Australian Research Council, Grant/Award Number: DE180100635

Abstract

Due to global climate change–induced shifts in species distributions, estimating changes in community composition through the use of Species Distribution Models has become a key management tool. Being able to determine how species associations change along environmental gradients is likely to be pivotal in exploring the magnitude of future changes in species' distributions. This is particularly important in connectivity-limited ecosystems, such as freshwater ecosystems, where increased human translocation is creating species associations over previously unseen environmental gradients. Here, we use a large-scale presence–absence dataset of freshwater fish from lakes across the Fennoscandian region in a Joint Species Distribution Model, to measure the effect of temperature on species associations. We identified a trend of negative associations between species tolerant of cold waters and those tolerant of warmer waters, as well as positive associations between several more warm-tolerant species, with these associations often shifting depending on local temperatures. Our results confirm that freshwater ecosystems can expect to see a large-scale shift towards communities dominated by more warm-tolerant species. While there remains much work to be done to predict exactly where and when local extinctions may take place, the model implemented provides a starting-point for the exploration of climate-driven community trends. This approach is especially informative in regards to determining which species associations are most central in shaping future community composition, and which areas are most vulnerable to local extinctions.

KEYWORDS
climate change, co-occurrence, fish, JSDMs

1 | INTRODUCTION

Due to global trends in species distribution range shifts and biodiversity loss, ecosystems worldwide are likely to undergo considerable changes in community composition (Jennings & Harris, 2017; Seebens et al., 2020). Along with the increased spread of non-native species as a result of globalisation, increasing average temperatures due to climate change will enable new species to make their way into environments previously too cold for them to either enter or establish in (Rahel & Olden, 2008; Sharma et al., 2007;

Walther et al., 2005). Many of these species are capable of causing extirpations of native species or even the restructuring of entire food webs once established (Nackley et al., 2017; Rockwell-Postel et al., 2020; Walther et al., 2009). Many species are vulnerable to local population declines and extinctions as increasing temperatures and extreme weather events compound threats posed by other anthropogenic factors, among them biological invasions (Dawson et al., 2011).

A changing climate is also capable of changing associations between species. While many species may be capable of co-occurring at certain temperatures, as annual temperatures increase, one species may gain a competitive advantage. For instance, at moderate temperatures a beech forest may contain three or four species, yet as temperatures tend to either extreme, negative associations may occur in the form of a single species beginning to outcompete the others and dominate (Leathwick, 2002). Associations becoming more negative with temperature increases could lead to declines in some species' populations, and local extinctions occurring, well before these species' upper thermal tolerances are reached. A deeper understanding of changes in species associations over a temperature gradient would contribute greatly to our understanding of the likely effects of climate change on community composition (Early & Keith, 2019; Freeman et al., 2018).

Climate change is likely to affect freshwater ecosystems particularly harshly, with a rise in temperature likely to lead to a population increase in species with higher thermal tolerance, with species of lower tolerance shifting further upstream (Comte et al., 2013; Daufresne & Boët, 2007). Increases in human translocations over recent decades are leading to novel species associations (Carpio et al., 2019), making understanding the impacts of these associations particularly important to predict future ecosystem effects. Research in sub-Arctic regions – which are likely to warm substantially in the coming decades – has already demonstrated a shift towards more warm-adapted species both within individual lakes and across catchments. (Hayden et al., 2017; Sharma et al., 2007; Van Zuiden et al., 2015; Winfield et al., 2008). Northern pike (*Esox Lucius*; Linnaeus, 1758) and brown trout (*Salmo trutta*; Linnaeus, 1758) may co-occur at lower temperatures in the sub-Arctic, but as average annual temperatures increase, a negative association results in a drop in the brown trout population as pike begin to predate brown trout at higher rates, eventually leading to local brown trout extinctions (Hein et al., 2013). The accelerated nature of climate change in the Arctic and sub-Arctic means that these regions are particularly important in giving an insight into potential community shifts, which are likely to take place in the coming decades in warmer regions.

Although much of the current research on shifts in associations over a temperature range concerns only a few species, or at a relatively small scale, attempts to increase focus on broader community models in predicting the impacts of climate change at a larger scale are ongoing (Comte et al., 2013; Radinger et al., 2019; Silknetter et al., 2020). Here, we construct a Joint Species Distribution Model (JSDM) as proposed by Tikhonov et al. (2017) to predict trends in species associations across a temperature gradient on a multi-national scale. JSDMs arose as a combination of habitat modelling and community ecology, in response to the need to account for associations which are not explained by the effects of environmental covariates (Ovaskainen et al., 2017; Pollock et al., 2014; Warton et al., 2015). Although they are capable of estimating correlations between species, these correlations may be a product of shared habitat specialisation, not interactions (Hargreaves et al., 2020; Hayden et al., 2013). As such, the output of JSDMs are often referred to as representing species associations, as opposed to interactions (Blanchet et al., 2020).

Here we use a dataset containing presence-absence data of fish species across more than 3000 freshwater lakes in the region of Fennoscandia, covering large climate gradients across mainland Norway, Sweden and Finland (Tammi et al., 2003). Using this modelling framework we aim to assess (a) whether we are able to identify the changes in associations between species of different thermal tolerances over a temperature gradient at large scales, (b) the reliability of our predictions based on whether these associations corroborate the results of previous research into pairwise associations and (c) how accurately these models can be used to predict future community shifts in the face of climate change. We aim to provide ecological modellers with a promising framework to build upon when exploring shifts in species distributions and abundances and provide managers with key information regarding potential extinction and invasion hotspots going forward.

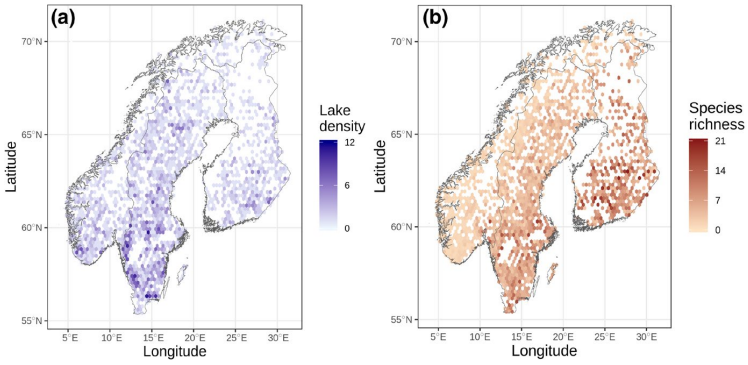


FIGURE 1 (a) Density of the 3308 lakes surveyed in the 1995 Nordic Fish Status Survey, represented by the number of lake centrepoints contained within hexagons. (b) Distribution of freshwater fish species richness across the same lakes, with species richness equal to average species richness of lakes found within hexagons

2 | METHODS

2.1 | Study system

Our study system was a series of 3308 lakes throughout the Fennoscandian region (Norway, Sweden and Finland). The lakes were located between 55.4 degrees and 71.1 degrees in latitude, 4.6 and 31.4 degrees longitude (WGS84), and at an altitudinal range of 0 to 1540 m. Surveyed lake density is highest throughout Sweden (Figure 1a).

The region itself is particularly species poor, given relatively recent deglaciation (Huitfeldt-Kaas, 1918). Species richness increases eastwards, with Finland and Sweden showing higher species richness than western Norway (Figure 1b). This is a product of mountainous regions dividing Norway, which have provided fewer immigration pathways for freshwater fish. Studies in the region have predicted that native species belonging to warmer guilds ('cool-water species'), like the Northern pike, the European perch (*Perca fluviatilis*; Linnaeus, 1758), and cyprinids like the common roach (*Rutilus rutilus*; Linnaeus, 1758) (Elliott, 2010; Hayden et al., 2014; Hokanson, 1977; Wehrly et al., 2003) will expand their range (Comte et al., 2013; Hayden et al., 2017). This could result in the local extirpation of species like the brown trout, whitefish (*Coregonus lavaretus*; Valenciennes, 1848) and Arctic charr (*Salvelinus alpinus*; Linnaeus, 1758), which are more tolerant of cold waters ('cool-cold/cold-water species'; Elliott, 2010; Hayden et al., 2014; Parkinson et al., 2016; Wehrly et al., 2003). The co-occurrence of species from various thermal guilds, within Fennoscandia, makes the region ideal for studying the impacts of climate change on freshwater communities (Comte et al., 2013).

2.2 | Occurrence data

The occurrence data were collected from a Fennoscandian survey of freshwater fish, originally conducted between 1995 and 1997 (Tammi et al., 2003). The resulting dataset consists of presence-absence information on fish species in 3821 lakes across Fennoscandia (Table 1). Henceforth, all species will be referred to by their common name (Table 1). Species that appeared in less than 1% of total lakes were excluded, as it is unlikely species-specific parameters can be accurately estimated given such low levels of occurrence. None were species thought likely to have a large-scale negative impact on Fennoscandian freshwater communities, either through competition or direct predation (for the full list of species, see Appendix S1).

2.3 | Environmental data

Temperature data were derived from the EuroLST data set for the centre of each lake (Metz et al., 2014). The temperature covariate used was the average air temperature of the warmest quarter, which has shown to have a strong correlation with water lake temperature (Livingstone & Lotter, 1998). This was the available data most likely to be strongly correlated with ice-off dates and spring warming temperatures, both of which have been shown to have a strong effect on both the life histories of aquatic species and interactions between different species (Mehner et al., 2011; Munsch et al., 2019).

For each lake, we obtained six additional covariates describing either environmental properties or human impact with the potential to influence establishment risk. Our study focuses on the effect of temperature; however, other environmental covariates are likely to have a strong effect on community composition, and as such are included as covariates. Environmental properties included lake surface area, shoreline complexity, total area of lakes situated upstream of focal lake, water pH, total organic carbon and human impact at the site. Area, shoreline complexity, and total upstream area were all derived from GIS analyses. Shoreline complexity was then calculated as:

$$SC = \frac{P}{2000\sqrt{\pi A}},$$

where P is the lake perimeter and A is the lake surface area (Wetzel, 2001). Water chemistry covariates were taken from the 1995 Nordic Lake Survey (Henriksen et al., 1998). To approximate human impact on each site, we used the Human Footprint Index as compiled by Venter at al. (2016), henceforth referred to as HFI. HFI is a point score which combines eight human impact covariates to approximate the level of human pressure on nature, assigned to cells one kilometre squared in size (Venter et al., 2016). These impact covariates include presence of built environment, crop lands and roads, and local human population density. HFI was taken for the cell in which the centre point of the lake lay, with previous research suggesting that human activity in the immediate vicinity of freshwater sites is more likely to affect species occurrences than activity upstream or downstream (Chapman et al., 2019). Lakes for which any of the environmental data was incomplete were disregarded (n = 119, 3.5% of total lakes). Area, total upstream area and total organic carbon were heavily right skewed and were thus log-transformed to assist with model convergence. All covariates were standardised by scaling to a mean of zero and a standard deviation of one to assist in model convergence. Further information on all covariates, included expected effects on populations, are summarised in Table 2.

The immigration history of freshwater fish in the area and steep topography that makes up much of Norway's west coast mean that many species have historically been unable to naturally colonise this region (Figure 1b; Huitfeldt-Kaas, 1918; Sandlund & Hesthagen, 2011). As such, environmental conditions which would normally result in the presence of species across other parts of Fennoscandia may have little to no effect on the likelihood of their presence in this region. This spatial divide could potentially make for a source of spatial autocorrelation, so to account for this we used the historical distribution range of species which were unable to colonise Norway's west coast as an additional covariate (see Appendix S2).

TABLE 1 Freshwater fish species surveyed in the 1995 Nordic Fish Status Survey

Common name	Scientific name	Family	Naming authority	Frequency of occurrence (%)
Perch	***Perca fluviatilis***	**Percidae**	**Linnaeus, 1754**	**72.1**
Pike	***Esox lucius***	**Esocidae**	**Linnaeus, 1754**	**65.8**
Roach	***Rutilus rutilus***	**Cyprinidae**	**Linnaeus, 1754**	**52.9**
Brown trout	***Salmo trutta***	**Salmonidae**	**Linnaeus, 1754**	**46.7**
Burbot	***Lota lota***	**Lotidae**	**Linnaeus, 1754**	**37.8**
Bream	***Abramis brama***	**Cyprinidae**	**Linnaeus, 1754**	**24.5**
Whitefish	***Coregonus lavaretus***	**Salmonidae**	**Valenciennes, 1844**	**23.5**
Ruffe	***Gymnocephalus cernuus***	**Percidae**	**Linnaeus, 1754**	**21.0**
Arctic charr	***Salvelinus alpinus***	**Salmonidae**	**Linnaeus, 1754**	**14.9**
Bleak	***Alburnus alburnus***	**Cyprinidae**	**Linnaeus, 1754**	**13.7**
Tench	***Tinca tinca***	**Cyprinidae**	**Linnaeus, 1754**	**13.5**
Vendace	***Coregonus albula***	**Salmonidae**	**Linnaeus, 1754**	**12.1**
Zander	***Stizostedion lucioperca***	**Percidae**	**Linnaeus, 1754**	**11.7**
Crucian carp	***Carassius carassius***	**Cyprinidae**	**Linnaeus, 1754**	**11.0**
Rudd	***Scardinius erythrophthalmus***	**Cyprinidae**	**Linnaeus, 1754**	**10.1**
Minnow	*Phoxinus phoxinus*	Cyprinidae	Linnaeus, 1754	9.2
Smelt	*Osmerus eperlanus*	Osmeridae	Linnaeus, 1754	8.1
White bream	*Blicca bjoerkna*	Cyprinidae	Linnaeus, 1754	6.5
Grayling	*Thymallus thymallus*	Salmonidae	Linnaeus, 1754	6.1
Ide	*Leuciscus idus*	Cyprinidae	Linnaeus, 1754	6.0
Rainbow trout	*Oncorhynchus mykiss*	Salmonidae	Walbaum, 1792	4.1
Threespine stickleback	*Gasterosteus aculeatus*	Gasterosteidae	Linnaeus, 1754	1.6
Brook trout	*Salvelinus fontinalis*	Salmonidae	Mitchill, 1814	1.4
Ninespine stickleback	*Pungitius pungitius*	Gasterosteidae	Linnaeus, 1754	1.0

Note: Table shows species taxonomy, as well as percentage of 3308 lakes that the species were found in. Species in bold occurred in more than 10% of lakes and were, thus, considered high-occurrence species.

2.4 | Statistical modelling

We constructed a JSDM which predicts changes in species associations over a temperature gradient. For our matrix of recorded observations, we assume that the presence-absences of species $j = 1...p$ at lakes $i = 1...n$ are independent observations, conditional on a vector of $h = 1...n_f$ latent factors per lake, modelled as:

$$y_{ij} \sim \text{Bernoulli}(p_{ij}),$$

$$\text{with } p_{ij} = \varphi^{-1}(\eta_{ij}),$$

where p_{ij} denotes the probability of species j being present at site i, and φ^{-1} denotes the inverse of a probit link-function. We denote η_{ij} as:

$$\eta_{ij} = \alpha_j + \sum_{k=1}^{n_c} x_{ik}\beta_{jk} + \varepsilon_{ij},$$

where x_{ik} denotes the value of environmental covariate $k = 1...n_c$ at site i, α_j denotes the intercept for species j, and β_{jk} denotes the effect of environmental covariate k on species j. The species-by-site random effects $\varepsilon_{ij} \sim N\left(0, R\left(x_i\right)\right)$ are defined by a latent factor model:

$$\varepsilon_{ij} = \sum_{h=1}^{n_f} z_{ih}\lambda_{jh}\left(x_i^*\right),$$

where z_{ih} denotes our $h = 1...n_f$ latent factors for lake i, where $\lambda_{jh}\left(x_i^*\right)$ denotes the responses (loading) of species $j = 1...p$ to each of the latent factors. We model the loadings per species j and for each factor h as a function of temperature:

$$\lambda_{jh}\left(x_i^*\right) = \sum_{i=1}^{n_l} \tau_{jh} + x_i u_{jh},$$

where τ_{jh} denotes an additional intercept for species j and latent factor h, u_{jh} denotes the response of latent factor h for species j to temperature, x_i denotes the temperature in degrees at site i, and n_l denotes the total number of sites. Here, we assume $n_f = 3$, as our Deviance information criteria did not improve significantly with the addition of more latent factors, and a few major gradients usually account for

TABLE 2 Environmental covariates, description, environmental effects, units and mean (±standard deviation) used in Joint Species Distribution Models of freshwater fish across European freshwater lakes

Environmental covariate	Description	Expected biological effect	Unit	Mean (±SD)
Area	Surface area of lake	Larger area increases potential habitat and niche breadth	Square kilometres	6163 (±52149)
Shoreline complexity	Calculated using area and perimeter	Increased shoreline complexity creates variation in habitat type (Verdiell-Cubedo et al., 2012)	Unitless	0.20, 0.14
Temperature	Average surface air temperature during maximum quarter	Temperature may alter various life history aspects of species (Magnuson et al., 1979)	Degrees	12.75, 1.74
Human Footprint Index	Index comprising 10 different variables, which represents impact of human activity (Venter at al., 2016)	Higher HFI increases chances of local human introductions (Chapman et al., 2019)	Unitless scale from 1 to 50	6.79, 7.03
Total upstream area	Aggregated area of lakes occurring directly upstream from focal lake	Higher upstream area increases chance of species' persistence	Square kilometres	388.68, 6405.17
pH	Taken from Nordic Lake Survey (Henriksen et al., 1998)	Acid sensitivity can limit local species' distributions (Ohman et al., 2006)	Unitless scale from 1 to 14	6.62, 0.66
Total organic carbon	Taken from Nordic Lake Survey (Henriksen et al., 1998)	Higher levels can cause anoxia and limit species' distributions (Ohman et al., 2006)	Mg per litre	7.09, 4.95
Biogeographic zone	Whether or not lake was found in a drainage basin cut-off from the rest of the region by the natural dispersal barrier running through central Norway	Presence of dispersal barrier provides fewer immigration pathways into western Norway for species, which did not colonise area via Norwegian Sea (Sandlund & Hesthagen, 2011)	Binary covariate	NA

most inter-species variation (Halvorsen, 2012). We define the matrices $\Lambda\left(x_i^*\right)$ with elements $\lambda_{jh}\left(x_i^*\right)$ and use these to construct a temperature-dependent, inter-species residual covariance matrix:

$$\Omega\left(x_i^*\right) = \Lambda\left(x_i^*\right)\Lambda\left(x_i^*\right)^T + I.$$

We then scale this covariance matrix to an inter-species correlation matrix R representing temperature-dependent associations between species that are not explained by fixed species-specific effects of environmental covariates:

$$R_{j_1j_2} = \Omega_{j_1j_2} / \sqrt{\Omega_{j_1j_1}\Omega_{j_2j_2}}.$$

These resulted in values between −1 and 1, with positive values indicative of positive associations between species, implying that species are likely to co-occur, and negative values implying the opposite. To compare species associations to the similarity in species responses to fixed effects in the models, we calculated an additional correlation matrix (Hui, 2017).

Although temperature was included in the random-effect, it was also included as a fixed-effect with quadratic function, to account for potential non-linear responses of species to temperature (Boddy & McIntosh, 2017; Veen et al., 2021).

We fit the model in a Bayesian framework using the greta R-package (Golding, 2019). All parameters were specified non-informative normally-distributed priors with a mean of zero and a standard deviation of 10, with the exception of the latent factors z_{ih} (mean = 0, SD = 1) and alpha parameters α_j (mean = −2, SD = 1). Alpha parameters had a lower mean to assist with convergence, as most species were prevalent at a low number of locations. Further exceptions were $\boldsymbol{u}$ and τ matrices, for which the diagonals had positively truncated non-informative normal priors, and all values in the upper triangle, which were set to zero to enforce identifiability (Hui et al., 2015). Markov Chain Monte-Carlo (MCMC) sampling was done

using 4000 samples on one chain, with a burn-in of 2000 samples. Although most Bayesian analyses would use multiple chains, latent variable models are often invariant to sign-switching (Hui, 2017), so here we choose to only run one chain. We used a Hamiltonian Monte Carlo sampler, sampling the number of leapfrog steps at each iteration uniformly between 40 and 60 (these numbers were manually tuned to achieve efficient sampling). The leapfrog integrator step sizes for each parameter were automatically tuned during the burn-in phase, then fixed for sampling. Parameters were considered to have adequately converged if their Geweke *Z*-score was below 1.96 (Geweke, 1992). 96 of our 105 species association parameters converged adequately. Trace plots for all association parameters can be found in Appendix S3.

We estimated changes in species associations over a continuous gradient from 6.68 to 16.80 degrees Celsius, which represented the minimum and maximum temperature observed in the data. Three models were constructed. Model 1 included data from all 24 species ("all species" model). Model 2 included data from species which only occurred in more than 10% of lakes (henceforth referred to as high-occurrence species, which are indicated as bold names in Table 1), as we wanted to test whether the inclusion of low-occurrence or low-detectability species produced a better model fit for commonly occurring species. Of the 24 species included in this study, 15 were classified as high-occurrence, and were thus used in model 2 (the "reduced species" model). To test whether accounting for species associations over a temperature gradient improved model fit, model 3 accounted for species associations, but not over a temperature gradient ("base JSDM" model). In this model, ε_{ij} is defined as:

$$\varepsilon_{ij} = \sum_{h=1}^{n_f} z_{ih}\lambda_{jh}.$$

Model fit was quantified using the Bernoulli deviance D_j, where

$$D_j = -2 \times \sum_i \left(y_{ij}\log\left(\mu_{ij}\right) + \left(1-y_{ij}\right)\log\left(1-\mu_{ij}\right)\right),$$

which was calculated for each high-occurrence species *j* of each model using the posterior medians of p_{ij}. To ensure that our models were an improvement over single species distribution models (SDMs), we created a stacked species distribution model (SSDM) consisting of single species distribution models for the fifteen species with occurrences in over 10% of lakes. These were also probit models with a Bernoulli distribution and used the same environmental covariates as our three previous models, but they did not include latent factors to account for associations between species. Model fit between the three models was compared using the improvement in deviance D_j from the SSDM for each of the high-occurrence species.

All statistical analyses were completed using R version 3.4.4 (R Core Team, 2017) and RStudio (RStudio Team, 2020). Additionally, the following R-packages were used for analysis and visualisation; dplyr (Wickham et al., 2019), rgeos (Bivand & Rundel, 2019), spdep (Bivand & Wong, 2018), postGIStools (Marchand & Ellison, 2019), tensorflow (Allaire & Tang, 2019), corrplot (Wei & Simko, 2017), ggplot2 (Wickham, 2016), magrittr (Bache & Wickham, 2014) and gridExtra (Auguie, 2017). A comprehensive definition of the model, the code, and its analysis can be found at Perrin (2021) (https://doi.org/10.5281/zenodo.4665778).

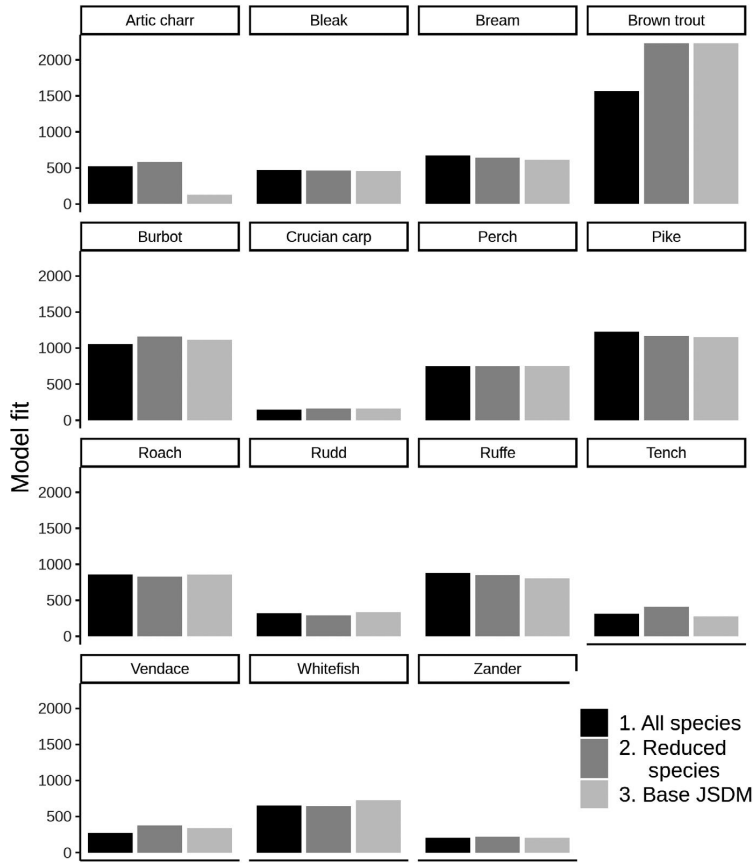
3 | RESULTS

Our results captured variations in species associations across a temperature gradient, with many negative associations between cold-water species and those tolerant of warmer temperatures.

Model fit was relatively similar across the all species model, reduced species model and base JSDM for each species (Figure 2). The only notable exceptions were brown trout, tench and arctic charr, for which the reduced species model (model two) had better model fit than either one of or both the other models. As such, the reduced species model was used for further analysis, as computing time was considerably lower. Biogeographic zone did not have a significant effect on any species and did not affect species associations, and as such the covariate was removed from all models.

Although several species associations at lower temperatures were negative – indicating a low likelihood of co-occurrence – at the mean and higher temperatures most associations between species were positive or close to neutral (Figure 3). The majority of negative associations between species at mean or high temperatures occurred between the cold-water species (brown trout or Arctic charr) and other species, with the most negative associations occurring between these species and those classified as belonging to a higher thermal guild (roach, perch and pike). Correlations in response to aggregated environmental variables were positive between the majority of species, with the exception of Arctic charr, for which many correlations were negative (Figure 4a). Correlations between species in responses to temperature were stronger than responses to all environmental variables, with a more even mix of positive and negative correlations (Figure 4b). Correlations in responses to each environmental covariate can be found in Appendix S4, as can species individual responses to environmental covariates.

As previous research has suggested potential associations between commonly occurring cool-water species (pike, perch, roach, whitefish) and cold-water species (brown trout and Arctic charr), these associations are shown in more detail in Figure 5, with thermal guild classifications found in Table 3. The association between brown trout and Arctic charr was negative at lower temperatures, becoming gradually positive before peaking at the mean temperature and decreasing as temperatures further increased. Arctic charr associations with perch, pike and roach were all negative at the lowest temperature, remained such until the mean temperature, and then increased slightly to be around neutral at higher temperatures. Arctic charr associations with whitefish increased consistently from being negative at the lowest temperature, before levelling out and remaining neutral at higher temperatures. Brown trout associations with perch, pike and roach all followed similar patterns, with



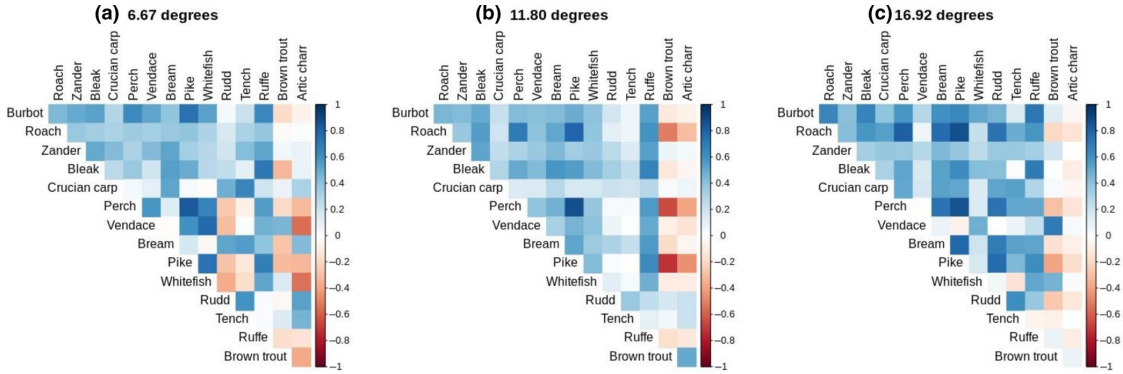


FIGURE 2 Model fit for three joint species distribution models mapping freshwater fish associations across 3308 Fennoscandian freshwater lakes. Model one utilises data from every species available, whereas model two utilises only data from species which occurred in over 10% of lakes. Models one and two estimate changes in species associations over a temperature gradient. Model three is a basic JSDM which monitors species associations, but does not estimate changes in associations over a temperature gradient. Model fit was measured using the improvement of each model's deviance values over the deviance values given by a stacked species distribution model which did not account for associations between species

FIGURE 3 Estimates of associations between 15 different freshwater fish species across freshwater lakes in 3308 lakes across the Fennoscandia region at different temperatures. Associations are shown for the region's (a) minimum (6.67 degrees), (b) mean (11.80 degrees) and (c) maximum (16.92 degrees) temperatures. Temperatures used represent average surface temperature during the warmest quarter of the year. Associations vary between 1 (most positive) and -1 (most negative)

associations close to neutral at the lowest temperature, becoming negative at the mean temperature and re-ascending towards zero as temperatures further increased. Associations between whitefish and brown trout were weakly positive at low temperatures, decreasing to weakly negative at the mean temperature before becoming more positive at higher temperatures. Whitefish associations with pike, perch and roach were positive at lower temperatures, and became weaker (although still positive) at higher temperatures. Associations

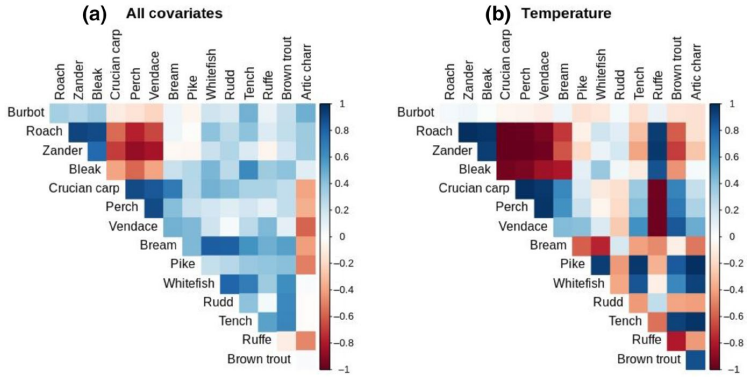


FIGURE 4 Correlation due to shared environmental response of 15 different fish species across 3308 freshwater lakes in Fennoscandia. Figures show correlation due to response to (a) all environmental covariates used in the Joint Species Distribution Model and (b) average surface air temperature of the warmest quarter of the year. Associations vary between 1 (most positive) and -1 (most negative)

between perch, pike and roach were all strongly positive, with little variation across temperature.

Maps visualising modelled predictions of likely changes in freshwater species distributions can be seen in Appendix S5.

4 | DISCUSSION

The ability to predict changes in species associations over environmental gradients will be crucial to incorporate into species distribution modelling as climate change modifies temperatures (Freeman et al., 2018). Here, we quantify changes in species associations over a climate gradient on a multi-national scale, using a presence-absence data set comprising 3308 lakes to fit a series of JSDMs of different freshwater fish species.

In accordance with Tikhonov et al. (2017) we demonstrate that large-scale presence-absence data are capable of shedding light on species associations over environmental gradients. Species belonging to cold-water guilds are generally negatively associated (and thus have a low likelihood of co-occurrence) with species from warmer thermal guilds, and these associations may change as temperatures warm.

Our confidence in these results is boosted by the fact that many of the associations predicted by our model corroborate previous research performed on pairwise associations. The negative associations between the two cold-water species and the cool-water species included here have been observed on smaller scales in this and similar study regions (Byström et al., 2007; Hayden et al., 2017; Hein et al., 2013; Winfield et al., 2008). Likewise, some of the positive associations shown among cool-water species here also have historical precedence (Eklöv & Hamrin, 1989; Mills & Hurley, 1990; Sharma & Borgstrøm, 2008).

A positive association between two species does not imply the lack of a negative impact of one species on another. Our results indicate a positive association between whitefish and perch, despite past evidence suggesting that whitefish are negatively impacted by the presence of perch (Hayden et al., 2013). However, it is possible for the two species to co-occur, e.g. through niche segregation (Hayden et al., 2014). However, since our response variable is binary, significant impacts on habitat use or life-history would not necessarily equate to a demonstrable negative impact in this study unless one species were driven to local extinction, unlike when fitting a latent variable model to abundance data.

It is important to note that predictions of species associations may become uninformative at certain temperatures. For example, when temperatures reach levels that preclude a species occurring in that region at all, any effect of species associations in an environmental context becomes void (Tikhonov et al., 2017). This is reflected in the associations between some species of different thermal guilds, which are predicted to increase towards zero as temperatures reach the higher ends of the spectrum. Summer temperatures in the region's warmer lakes are higher than the temperature range of lakes typically occupied by cold-water species (Mandeville et al., 2019). As such our predictions of associations between species should only be considered reliable at temperature ranges where both species are capable of persisting independently.

Although such models are capable of estimating future shifts in community composition in response to climate change, we recommend instead treating estimates provided by such models as indications of potential larger trends and – similar to Wagner at al. (2020) – as a basis for generating hypotheses and focussing future research (Zurell et al., 2020). Although some of the associations here match previous research, others point to new potential threats to native cold-water species like the Arctic charr and brown trout, which should be studied more thoroughly. Although it was beyond the scope of this paper, further research could also take into account possible interactions between temperature and other environmental covariates, for instance habitat area, as research has suggested that often colder-tolerant species can withstand potential competition if there is enough available habitat and niche segregation within a given habitat patch (Hein et al., 2013).

These results show that on a broad, multi-national scale, shifts towards communities dominated by species which have higher thermal tolerance are likely to occur as climate change drives average temperatures higher. Although many lakes may not reach the thermal maximum of native species, our model confirms that local extinctions are likely to occur earlier, driven by changing associations between native species and either invasive non-native species,

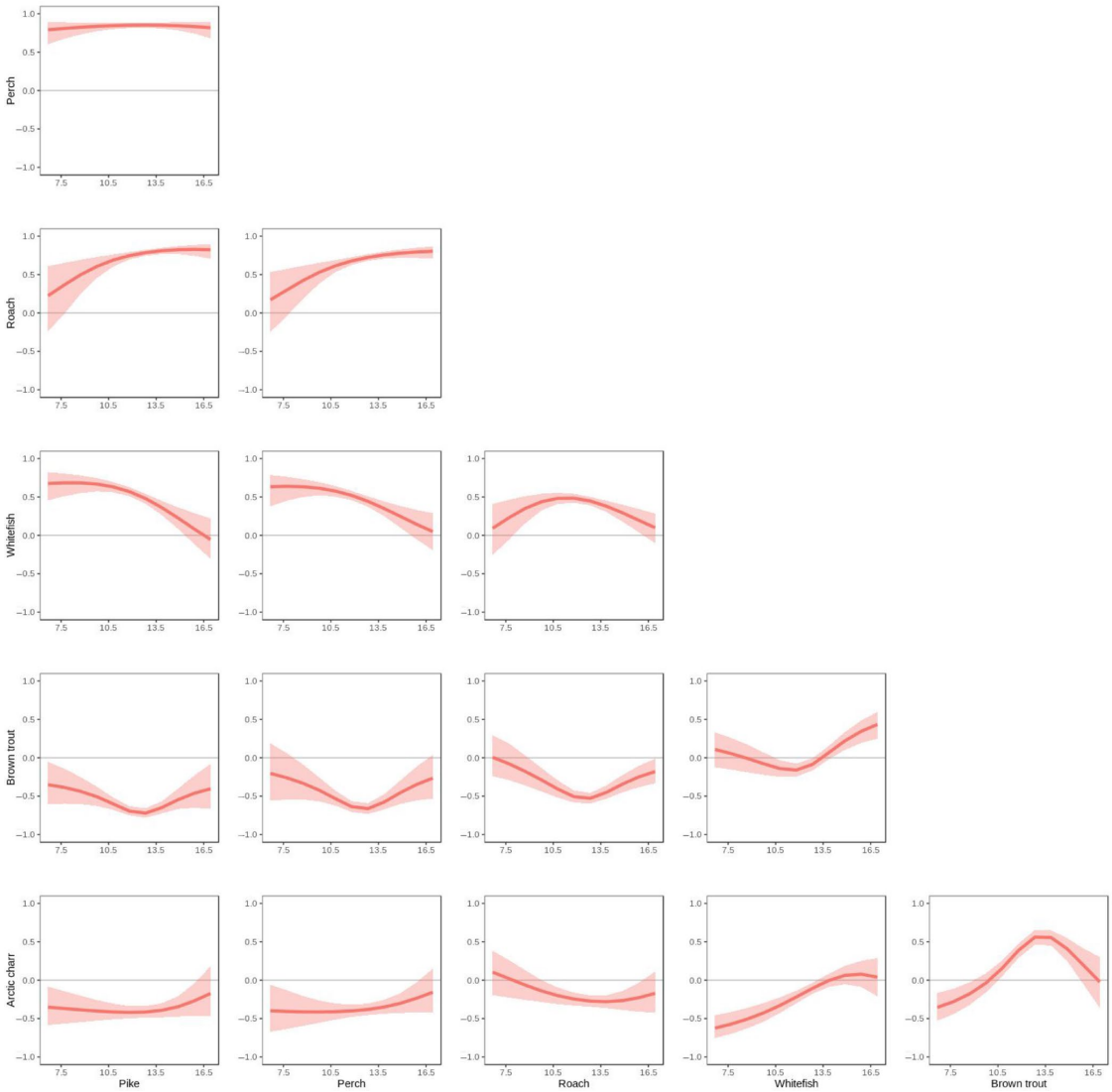


FIGURE 5 Associations between six freshwater fish species over a temperature gradient inferred from residual correlations from a joint species distribution model across 3308 freshwater lakes in the Fennoscandian region. Temperatures displayed on *x*-axes represent average surface temperature during the warmest quarter of the year. Associations are displayed on *y*-axes on a scale from −1 to 1, with ribbons representing 95% credible intervals. Negative values represent negative associations between species

TABLE 3 Thermal guild classification of six freshwater fish species surveyed in 1995 Nordic Fish Status Survey, with references citing precedent for classification

Common name	Thermal guild	References
Perch	Cool	Hayden et al. (2014); Hokanson (1977)
Pike	Cool	Wehrly et al. (2003)
Roach	Cool	Elliott (2010)
Whitefish	Cool-cold	Hayden et al. (2014)
Brown trout	Cold	Elliott (2010); Wehrly et al. (2003)
Arctic charr	Cold	Elliott (2010)

range shifting species or species with which native fish had previously co-occurred. The tendency of many species towards positive associations with an increase in temperature suggests a trend towards homogenisation of freshwater communities, though a variation in species individual responses to increased temperature could affect this.

Previous research indicates that more cold-tolerant species are capable of persisting in larger, deeper lakes due to the possibility of spatial segregation (Hein et al., 2013). Areas with strong topographical variation could provide dispersal barriers for novel species, and subsequently provide refugia for species likely to be outcompeted (Perrin et al., 2020). However such refugia are only likely to be tenable if human translocation is sufficiently regulated so as to prevent the introduction of novel species (Hesthagen & Sandlund, 2004; Perrin et al., 2021). As such, models like the one constructed here could enable researchers not only to identify environmental covariates which may drive changes in species associations and subsequently identify areas where native species are vulnerable to local extinctions, but also to identify areas where such species are likely to persist.

The ability of large-scale SDMs to predict changes in species associations while corroborating smaller-scale pairwise research over a temperature gradient is encouraging. It implies that large-scale presence-absence data may be capable of predicting changes in community composition as temperatures increase in the coming decades. Although much work remains to ensure the accuracy and reliable management application of such models, our results here indicate that JSDMs can be used to identify the potential impacts of climate change and range-shifting species on global ecosystems.

ACKNOWLEDGEMENTS

The authors would like to thank Göran Englund for advancing their understanding of the 1995 Nordic Freshwater Fish Survey and Bob O'Hara and Keller Kopf for helpful comments on an earlier version of the manuscript and initial project design. Sam Perrin was supported by a PhD grant from the ERA-Net BiodivERsA project ODYSSEUS (Norwegian Research Council 266574). Nick Golding was supported by an Australian Research Council Discovery Early Career Researcher Award (DE180100635).

CONFLICT OF INTEREST

No conflict of interest for the article.

DATA AVAILABILITY STATEMENT

Data and code used for the species distribution modelling are archived in Zenodo at http://doi.org/10.5281/zenodo.4665778.

ORCID

Sam Wenaas Perrin https://orcid.org/0000-0002-1266-1573
Bert van der Veen https://orcid.org/0000-0003-2263-3880
Nick Golding https://orcid.org/0000-0001-8916-5570
Anders Gravbrøt Finstad https://orcid.org/0000-0003-4529-6266

REFERENCES

Allaire, J. J., & Tang, Y. (2019). tensorflow: R interface to "TensorFlow". https://github.com/rstudio/tensorflow

Auguie, B. (2017). gridExtra: Miscellaneous functions for "grid" graphics. https://CRAN.R-project.org/package=gridExtra

Bache, S. M. & Wickham, H. (2014). magrittr: A forward-pipe operator for R. https://CRAN.R-project.org/package=magrittr

Bivand, R. & Rundel, C. (2019). rgeos: Interface to Geometry Engine - Open Source ('GEOS'). https://CRAN.R-project.org/package=rgeos

Bivand, R. S., & Wong, D. W. S. (2018). Comparing implementations of global and local indicators of spatial association. *TEST*, *27*, 716–748. https://doi.org/10.1007/s11749-018-0599-x

Blanchet, F. G., Cazelles, K., & Gravel, D. (2020). Co-occurrence is not evidence of ecological interactions. *Ecology Letters*, *23*, 1050–1063. https://doi.org/10.1111/ele.13525

Boddy, N. C., & McIntosh, A. R. (2017). Temperature, invaders and patchy habitat interact to limit the distribution of a vulnerable freshwater fish. *Austral Ecology*, *42*(4), 456–467. https://doi.org/10.1111/aec.12463

Byström, P., Karlsson, J., Nilsson, P., Van Kooten, T., Ask, J., & Olofsson, F. (2007). Substitution of top predators: Effects of pike invasion in a subarctic lake. *Freshwater Biology*, *52*(7), 1271–1280. https://doi.org/10.1111/j.1365-2427.2007.01763.x

Carpio, A. J., De Miguel, R. J., Oteros, J., Hillström, L., & Tortosa, F. S. (2019). Angling as a source of non-native freshwater fish: A European review. *Biological Invasions*, *5*, 3233–3248. https://doi.org/10.1007/s10530-019-02042-5

Chapman, D. S., Gunn, I. D. M., Pringle, H. E. K., Siriwardena, G. M., Taylor, P., Thackeray, S. J., Willby, N. J., & Carvalho, L. (2019). Invasion of freshwater ecosystems is promoted by network connectivity to hotspots of human activity. *Global Ecology and Biogeography: A Journal of Macroecology*, *88*, 528.

Comte, L., Buisson, L., Daufresne, M., & Grenouillet, G. (2013). Climate-induced changes in the distribution of freshwater fish: Observed and predicted trends. *Freshwater Biology*, *58*(4), 625–639. https://doi.org/10.1111/fwb.12081

Daufresne, M., & Boët, P. (2007). Climate change impacts on structure and diversity of fish communities in rivers. *Global Change Biology*, *13*(12), 2467–2478. https://doi.org/10.1111/j.1365-2486.2007.01449.x

Dawson, T. P., Jackson, S. T., House, J. I., Prentice, I. C., & Mace, G. M. (2011). Beyond predictions: Biodiversity conservation in a changing climate. *Science*, *332*(6025), 53–58.

Early, R., & Keith, S. A. (2019). Geographically variable biotic interactions and implications for species ranges. *Global Ecology and Biogeography: A Journal of Macroecology*, *28*(1), 42–53. https://doi.org/10.1111/geb.12861

Eklöv, P., Hamrin, S. F., & Eklov, P. (1989). Predatory efficiency and prey selection: Interactions between pike *Esox lucius*, perch *Perca fluviatilis* and rudd *Scardinus erythrophthalmus*. *Oikos*, *56*(2), 149–156. https://doi.org/10.2307/3565330

Elliott, A. (2010). A comparison of thermal polygons for British freshwater teleosts. *Freshwater Forum*, *5*. https://core.ac.uk/download/pdf/228601267.pdf

Freeman, B. G., Lee-Yaw, J. A., Sunday, J. M., & Hargreaves, A. L. (2018). Expanding, shifting and shrinking: The impact of global warming on species' elevational distributions. *Global Ecology and Biogeography: A Journal of Macroecology*, *27*(11), 1268–1276.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments. *Bayesian Statistics*, *4*, 641–649.

Golding, N. (2019). greta: Simple and scalable statistical modelling in R. *Journal of Open Source Software*, *4*(40), 1601. https://doi.org/10.21105/joss.01601

Halvorsen, R. (2012). A gradient analytic perspective on distribution modelling. *Sommerfeltia*, *35*(1), 1–165. https://doi.org/10.2478/v10208-011-0015-3

Hargreaves, A. L., Germain, R. M., Bontrager, M., Persi, J., & Angert, A. L. (2020). Local adaptation to biotic interactions: A meta-analysis across latitudes. *The American Naturalist*, *195*(3), 395–411. https://doi.org/10.1086/707323

Hayden, B., Harrod, C., & Kahilainen, K. K. (2014). Lake morphometry and resource polymorphism determine niche segregation between cool-and cold-water-adapted fish. *Ecology*, *95*(2), 538–552. https://doi.org/10.1890/13-0264.1

Hayden, B., Holopainen, T., Amundsen, P. A., Eloranta, A. P., Knudsen, R., Præbel, K., & Kahilainen, K. K. (2013). Interactions between invading benthivorous fish and native whitefish in subarctic lakes. *Freshwater Biology*, *58*(6), 1234–1250. https://doi.org/10.1111/fwb.12123

Hayden, B., Myllykangas, J. P., Rolls, R. J., & Kahilainen, K. K. (2017). Climate and productivity shape fish and invertebrate community structure in subarctic lakes. *Freshwater Biology*, *62*(6), 990–1003. https://doi.org/10.1111/fwb.12919

Hein, C. L., Ohlund, G., & Englund, G. (2013). Fish introductions reveal the temperature dependence of species interactions. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1775), 20132641.

Henriksen, A., Skjelvåle, B. L., Mannio, J., Wilander, A., Harriman, R., Curtis, C., Jensen, J. P., Fjeld, E., & Moiseenko, T. (1998). Northern European Lake Survey, 1995: Finland, Norway, Sweden, Denmark, Russian Kola, Russian Karelia, Scotland and Wales. *Ambio*, *27*(2), 80–91.

Hesthagen, T., & Sandlund, O. T. (2004). Fish distribution in a mountain area in south-eastern Norway: Human introductions overrule natural immigration. *Hydrobiologia*, *521*, 49–59. https://doi.org/10.1023/B:HYDR.0000026350.93171.ba

Hokanson, K. E. F. (1977). Temperature requirements of some percids and adaptations to the seasonal temperature cycle. *Journal of the Fisheries Research Board of Canada*, *34*(10), 1524–1550. https://doi.org/10.1139/f77-217

Hui, F. K. C. (2017). Boral: Bayesian ordination and regression analysis. R Package Version, 1.

Hui, F. K. C., Taskinen, S., Pledger, S., Foster, S. D., & Warton, D. I. (2015). Model-based approaches to unconstrained ordination. *Methods in Ecology and Evolution*, *6*(4), 399–411. https://doi.org/10.1111/2041-210X.12236

Huitfeldt-Kaas, H. (1918). *Ferskvandsfiskenes utbredelse og indvandring i Norge, med et tillæg om krebsen*. Centraltrykkeriet.

Jennings, M. D., & Harris, G. M. (2017). Climate change and ecosystem composition across large landscapes. *Landscape Ecology*, *32*(1), 195–207. https://doi.org/10.1007/s10980-016-0435-1

Leathwick, J. R. (2002). Intra-generic competition among Nothofagus in New Zealand's primary indigenous forests. *Biodiversity & Conservation*, *11*(12), 2177–2187.

Livingstone, D. M., & Lotter, A. F. (1998). The relationship between air and water temperatures in the lakes of the Swiss Plateau: A case study with paleolimnological implications. *Journal of Paleolimnology*, *19*(2), 181–198.

Magnuson, J. J., Crowder, L. B., & Medvick, P. A. (1979). Temperature as an ecological resource. *American Zoologist*, *19*(1), 331–343. https://doi.org/10.1093/icb/19.1.331

Mandeville, C. P., Rahel, F. J., Patterson, L. S., & Walters, A. W. (2019). Integrating fish assemblage data, modeled stream temperatures, and thermal tolerance metrics to develop thermal guilds for water temperature regulation: Wyoming case study. *Transactions of the American Fisheries Society*, *148*(4), 739–754. https://doi.org/10.1002/tafs.10169

Marchand, P. & Ellison, R. (2019). postGIStools: Tools for interacting with "PostgreSQL" / "PostGIS" databases. https://CRAN.R-project.org/package=postGIStools

Mehner, T., Emmrich, M., & Kasprzak, P. (2011). Discrete thermal windows cause opposite response of sympatric cold-water fish species to annual temperature variability. *Ecosphere*, *2*(9), 1–16. https://doi.org/10.1890/ES11-00109.1

Metz, M., Rocchini, D., & Neteler, M. (2014). Surface temperatures at the continental scale: Tracking changes with remote sensing at unprecedented detail. *Remote Sensing*, *6*(5), 3822–3840; https://doi.org/10.3390/rs6053822

Mills, C. A., & Hurley, M. A. (1990). Long-term studies on the Windermere populations of perch (*Perca fluviatilis*), pike (*Esox lucius*) and Arctic charr (*Salvelinus alpinus*). *Freshwater Biology*, *23*(1), 119–136. https://doi.org/10.1111/j.1365-2427.1990.tb00257.x

Munsch, S. H., Greene, C. M., Johnson, R. C., Satterthwaite, W. H., Imaki, H., & Brandes, P. L. (2019). Warm, dry winters truncate timing and size distribution of seaward-migrating salmon across a large, regulated watershed. *Ecological Applications*, *29*(4), e01880. https://doi.org/10.1002/eap.1880

Nackley, L. L., West, A. G., Skowno, A. L., & Bond, W. J. (2017). The nebulous ecology of native invasions. *Trends in Ecology & Evolution*, *32*(11), 814–824. https://doi.org/10.1016/j.tree.2017.08.003

Ohman, J., Buffam, I., Englund, G., Blom, A., Lindgren, E., & Laudon, H. (2006). Associations between water chemistry and fish community composition: A comparison between isolated and connected lakes in northern Sweden. *Freshwater Biology*, *51*(3), 510–522. https://doi.org/10.1111/j.1365-2427.2006.01514.x

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T., & 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. https://doi.org/10.1111/ele.12757

Parkinson, E. A., Lea, E. V., Nelitz, M. A., Knudson, J. M., & Moore, R. D. (2016). Identifying temperature thresholds associated with fish community changes in British Columbia, Canada, to support identification of temperature sensitive streams. *River Research and Applications*, *32*(3), 330–347. https://doi.org/10.1002/rra.2867

Perrin, S. W. (2021). *samaperrin/ChangingAssociations* (Version v2) [Computer software]. https://doi.org/10.5281/zenodo.4665778

Perrin, S. W., Bærum, K. M., Helland, I. P., & Finstad, A. G. (2021). Forecasting the future establishment of invasive alien freshwater fish species. *Journal of Applied Ecology*. https://doi.org/10.1111/1365-2664.13993

Perrin, S. W., Englund, G., Blumentrath, S., O'Hara, R. B., Amundsen, P.-A., & Finstad, A. G. (2020). Integrating dispersal along freshwater ecosystems into species distribution models. *Diversity and Distributions*, *26*(11), 1598–1611. https://doi.org/10.1111/ddi.13112

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

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-Project.org/

Radinger, J., Alcaraz-Hernández, J. D., & García-Berthou, E. (2019). Environmental filtering governs the spatial distribution of alien fishes in a large, human-impacted Mediterranean river. *Diversity & Distributions*, *25*(5), 701–714. https://doi.org/10.1111/ddi.12895

Rahel, F. J., & Olden, J. D. (2008). Assessing the effects of climate change on aquatic invasive species source. *Conservation Biology*, *22*(3), 521–533. https://doi.org/10.1111/j.1523-1739.2008.00950.x

Rockwell-Postel, M., Laginhas, B. B., & Bradley, B. A. (2020). Supporting proactive management in the context of climate change: Prioritizing range-shifting invasive plants based on impact. *Biological Invasions*, *22*(7), 2371–2383. https://doi.org/10.1007/s10530-020-02261-1

RStudio Team. (2020). *RStudio: Integrated development environment for R*. RStudio, Inc. http://www.rstudio.com/

Sandlund, O. T., & Hesthagen, T. (2011). Fish diversity in Norwegian lakes: Conserving species poor systems. In M. Jankun, G. Furghala-Selezniow, M. Wozniak, & A. M. Wisniewska (Eds.), *Water biodiversity assessment and protection* (pp. 7–20). University of Warmia and Mazury.

Seebens, H., Bacher, S., Blackburn, T. M., Capinha, C., Dawson, W., Dullinger, S., Genovesi, P., Hulme, P. E., van Kleunen, M., Kühn, I., Jeschke, J. M., Lenzner, B., Liebhold, A. M., Pattison, Z., Pergl, J., Pyšek, P., Winter, M., & Essl, F. (2020). Projecting the continental accumulation of alien species through to 2050. *Global Change Biology*, *27*(5), 970–982. https://doi.org/10.1111/gcb.15333

Sharma, C. M., & Borgstrøm, R. (2008). Shift in density, habitat use, and diet of perch and roach: An effect of changed predation pressure after manipulation of pike. *Fisheries Research*, *91*(1), 98–106. https://doi.org/10.1016/j.fishres.2007.11.011

Sharma, S., Jackson, D. A., Minns, C. K., & Shuter, B. J. (2007). Will northern fish populations be in hot water because of climate change? *Global Change Biology*, *13*, 2052–2064. https://doi.org/10.1111/j.1365-2486.2007.01426.x

Silknetter, S., Creed, R. P., Brown, B. L., Frimpong, E. A., Skelton, J., & Peoples, B. K. (2020). Positive biotic interactions in freshwaters: A review and research directive. *Freshwater Biology*, *65*(4), 811–832. https://doi.org/10.1111/fwb.13476

Tammi, J., Appelberg, M., Beier, U., Hesthagen, T., Lappalainen, A., & Rask, M. (2003). Fish status survey of Nordic lakes: Effects of acidification, eutrophication and stocking activity on present fish species composition. *Ambio*, *32*(2), 98–105. https://doi.org/10.1579/0044-7447-32.2.98

Tikhonov, G., Abrego, N., Dunson, D., & Ovaskainen, O. (2017). Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context. *Methods in Ecology and Evolution*, *8*(4), 443–452. https://doi.org/10.1111/2041-210X.12723

Van der Veen, B., Hui, F. K. C., Hovstad, K. A., Solbu, E. B., & O'Hara, R. B. (2021). Model-based ordination for species with unequal niche widths. *Methods in Ecology and Evolution*, *12*(7), 1288–1300. https://doi.org/10.1111/2041-210X.13595

Van Zuiden, T. M., Chen, M. M., Stefanoff, S., Lopez, L., & Sharma, S. (2015). Projected impacts of climate change on three freshwater fishes and potential novel competitive interactions. *Diversity and Distributions*, *22*(5), 603–614. https://doi.org/10.1111/ddi.12422

Venter, O., Sanderson, E. W., Magrach, A., Allan, J. R., Beher, J., Jones, K. R., Possingham, H. P., Laurance, W. F., Wood, P., Fekete, B. M., Levy, M. A., & Watson, J. E. M. (2016). Global terrestrial Human Footprint maps for 1993 and 2009. *Scientific Data*, *3*, 160067. https://doi.org/10.1038/sdata.2016.67

Verdiell-Cubedo, D., Torralva, M., Andreu-Soler, A., & Oliva-Paterna, F. J. (2012). Effects of shoreline urban modification on habitat structure and fish community in littoral areas of a Mediterranean Coastal Lagoon (Mar Menor, Spain). *Wetlands*, *32*(4), 631–641. https://doi.org/10.1007/s13157-012-0296-6

Wagner, T., Hansen, G. J. A., Schliep, E. M., Bethke, B. J., Honsey, A. E., Jacobson, P. C., Kline, B. C., & White, S. L. (2020). Improved understanding and prediction of freshwater fish communities through the use of joint species distribution models. *Canadian Journal of Fisheries and Aquatic Sciences*, *77*(9), 1540–1551. https://doi.org/10.1139/cjfas-2019-0348

Walther, G.-R., Beisner, S., & Burga, C. A. (2005). Trends in the upward shift of alpine plants. *Journal of Vegetation Science*, *16*(1998), 541–548.

Walther, G.-R., Roques, A., Hulme, P. E., Sykes, M. T., Pysek, P., Kühn, I., Zobel, M., Bacher, S., Botta-Dukát, Z., Bugmann, H., Czúcz, B., Dauber, J., Hickler, T., Jarosík, V., Kenis, M., Klotz, S., Minchin, D., Moora, M., Nentwig, W., … Settele, J. (2009). Alien species in a warmer world: Risks and opportunities. *Trends in Ecology & Evolution*, *24*(12), 686–693. https://doi.org/10.1016/j.tree.2009.06.008

Warton, D. I., Blanchet, F. G., O'Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C., & Hui, F. K. C. (2015). So many variables: Joint modeling in community ecology. *Trends in Ecology & Evolution*, *30*(12), 766–779. https://doi.org/10.1016/j.tree.2015.09.007

Wehrly, K. E., Wiley, M. J., & Seelbach, P. W. (2003). Classifying regional variation in thermal regime based on stream fish community patterns. *Transactions of the American Fisheries Society*, *132*(1), 18–38.

Wei, T., & Simko, V. (2017). R package "corrplot": Visualization of a correlation matrix. https://github.com/taiyun/corrplot

Wetzel, R. G. (2001). *Limnology: Lake and river ecosystems*. Gulf Professional Publishing.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer.

Wickham, H., François, R., Henry, L., & Müller, K. (2019). dplyr: A grammar of data manipulation. https://CRAN.R-project.org/package=dplyr

Winfield, I. J., Fletcher, J. M., & James, J. B. (2008). The Arctic charr (*Salvelinus alpinus*) populations of Windermere, UK: Population trends associated with eutrophication, climate change and increased abundance of roach (*Rutilus rutilus*). *Environmental Biology of Fishes*, *83*(1), 25–35. https://doi.org/10.1007/s10641-007-9235-4

Zurell, D., Zimmermann, N. E., Gross, H., Baltensweiler, A., Sattler, T., & Wüest, R. O. (2020). Testing species assemblage predictions from stacked and joint species distribution models. *Journal of Biogeography*, *47*(1), 101–113. https://doi.org/10.1111/jbi.13608

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Perrin, S. W., van der Veen, B., Golding, N., & Finstad, A. G. (2021). Modelling temperature-driven changes in species associations across freshwater communities. *Global Change Biology*, 00, 1–12. https://doi.org/10.1111/gcb.15888

Supplementary Information Table S1: Freshwater fish species surveyed in 1995 Nordic Fish Status Survey. Table shows species taxonomy and naming authority, as well as percentage of 3308 lakes that the species were found in. Species for which frequency of occurrence is less than 1% were not used in species distribution modelling.

Common name	Scientific name	Family	Frequency of occurrence (%)	Naming Authority
Perch	Perca fluviatilis	Percidae	72.1	Linnaeus, 1754
Pike	Esox lucius	Esocidae	65.8	Linnaeus, 1754
Roach	Rutilus rutilus	Cyprinidae	52.9	Linnaeus, 1754
Brown trout	Salmo trutta	Salmonidae	46.7	Linnaeus, 1754
Burbot	Lota lota	Lotidae	37.8	Linnaeus, 1754
Bream	Abramis brama	Cyprinidae	24.5	Linnaeus, 1754
Whitefish	Coregonus lavaretus	Salmonidae	23.5	Valenciennes, 1844
Ruffe	Gymnocephalus cernuus	Percidae	21.0	Linnaeus, 1754
Arctic char	Salvelinus alpinus	Salmonidae	14.9	Linnaeus, 1754
Bleak	Alburnus alburnus	Cyprinidae	13.7	Linnaeus, 1754
Tench	Tinca tinca	Cyprinidae	13.5	Linnaeus, 1754
Vendace	Coregonus albula	Salmonidae	12.1	Linnaeus, 1754
Pike-perch	Stizostedion lucioperca	Percidae	11.7	Linnaeus, 1754
Crucian carp	Carassius carassius	Cyprinidae	11.0	Linnaeus, 1754
Rudd	Scardinius erythrophthalmus	Cyprinidae	10.1	Linnaeus, 1754
Minnow	Phoxinus phoxinus	Cyprinidae	9.2	Linnaeus, 1754
Smelt	Osmerus eperlanus	Osmeridae	8.1	Linnaeus, 1754
White bream	Blicca bjoerkna	Cyprinidae	6.5	Linnaeus, 1754
Grayling	Thymallus thymallus	Salmonidae	6.1	Linnaeus, 1754
Ide	Leuciscus idus	Cyprinidae	6.0	Linnaeus, 1754
Rainbow trout	Oncorhynchus mykiss	Salmonidae	4.1	Walbaum, 1792
Threespine stickleback	Gasterosteus aculeatus	Gasterosteidae	1.6	Linnaeus, 1754
Brook trout	Salvelinus fontinalis	Salmonidae	1.4	Mitchill, 1814
Ninespine stickleback	Pungitius pungitius	Gasterosteidae	1.0	Linnaeus, 1754
European	Lampetra planeri	Petromyzontidae	<1.0	Bloch, 1784

lamprey				
Gudgeon	Gobio gobio	Cyprinidae	<1.0	Linnaeus, 1754
Danube catfish	Silurus glanis	Siluridae	<1.0	Linnaeus, 1754
Baltic vimba	Vimba vimba	Cyprinidae	<1.0	Linnaeus, 1754
Common dace	Leuciscus leuciscus	Cyprinidae	<1.0	Linnaeus, 1754
Spine loach	Cobitis taenia	Cobitidae	<1.0	Linnaeus, 1754
Common carp	Cyprinus carpio	Cyprinidae	<1.0	Linnaeus, 1754
Chub	Leuciscus cephalus	Cyprinidae	<1.0	Linnaeus, 1754
NA	Leucaspius cephalus	Cyprinidae	<1.0	Heckel & Kner. 1858
European river lamprey	Lampetra fluviatilis	Petromyzontidae	<1.0	Linnaeus, 1754
Fourhorn sculpin	Myoxocephalus quadricornis	Cottidae	<1.0	Linnaeus, 1754
Atlantic salmon	Salmo salar	Salmonidae	<1.0	Linnaeus, 1754
Lake trout	Salvelinus namaycush	Salmonidae	<1.0	Walbaum, 1792
NA	Pelecus culturas	Cyprinidae	<1.0	Agassiz, 1835
Asp	Aspius aspius	Cyprinidae	<1.0	Linnaeus, 1754
Alpine bullhead	Cottus poecilopus	Cottidae	<1.0	Heckel, 1840
Grass carp	Ctenopharyngodon idella	Cyprinidae	<1.0	Valenciennes, 1844

Supplementary Figures S2 for *Estimating community-level changes in freshwater species associations over a temperature gradient*

The immigration history of freshwater fish into Fennoscandia, combined with the steep topography that makes up much of Norway's west coast mean that many freshwater fish species have historically been unable to naturally colonise this region (Huitfeldt-Kaas, 1918; Sandlund & Hesthagen, 2011). This has resulted in a stark spatial divide in species distributions across Fennoscandia, with much of Norway's west coast only having two or three native species. As such, environmental conditions which would normally result in the presence of species across other parts of Fennoscandia may have little to no effect on the likelihood of their presence in this region. The exceptions are anadromous species which colonised the west coast of Norway via the North Sea.

This spatial pattern is a potential source of spatial autocorrelation, with lakes within the same spatial boundaries - here interconnected groups of lakes found within the same water drainage basin - more likely to show similar assemblages. Distribution of lakes among water drainage basins was highly uneven, which prevented the use of drainage basins as a useful covariate. As such, we used the historical ranges of species which were unable to colonise the west coast of Norway to approximate the spatial divide across the region.

Species historical ranges were estimated using surveys originally published by Huitfelt-Kaas (1918), and later digitised and georeferenced (Daverdin et al. 2019). The historical ranges were merged into one large distribution polygon using the sf R-package (Pebesma, 2018), intended to represent the limitations of dispersal for species unable to colonise Norway via the west coast. Drainage basins which did not

intercept this polygon were assumed to be disconnected from the rest of the region by the natural dispersal barrier. These drainage basins were explicitly found along the west coast of Norway. The resulting binary covariate was built based on whether or not a lake was found inside these drainage basins. This covariate, termed the biogeographic zone, used in the model.

As our response variable took the form of discrete values, we used randomised quantile residuals to test for any residual spatial autocorrelation as defined by Bunn & Smyth (1996). Examples of spatial distribution of the residuals can be seen in Figure S2.1. A comprehensive description of the code used can be found in Perrin et al., 2021 (DOI: 10.5281/zenodo.4665778).

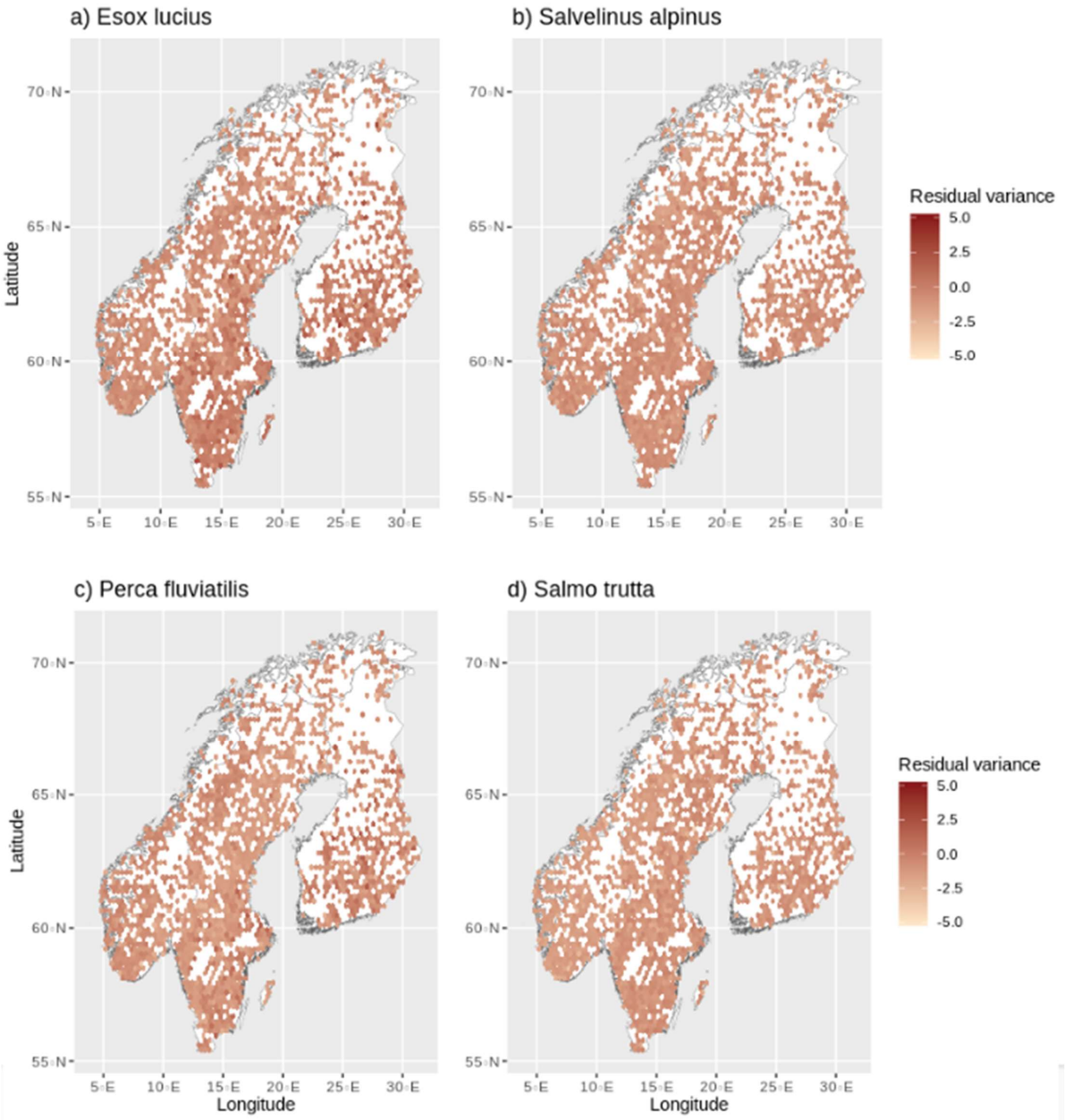


Figure S2.1: Spatial variation in residual variance of large-scale Joint Species Distribution Modelling of a) pike, b) Arctic charr, c) perch and d) brown trout. The residuals were Randomized Quantile Residuals (Dunn and Smyth, 1996)

References

Daverdin, M., Finstad, A. G., & Blumentrath, S. (2019) *Freshwater fish native distribution map transcriptions from Huitfeldt-Kaas, H. (1918)*. https://doi.org/10.21400/1MWT3950

Dunn, P. K., & Smyth, G. K. (1996). Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics: A Joint Publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, *5*(3), 236–244.

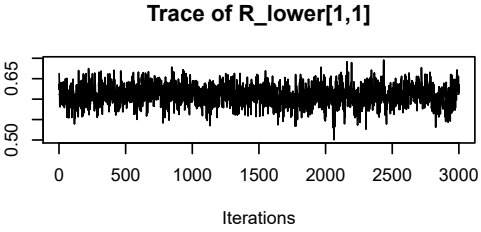
Huitfeldt-Kaas, H. (1918). *Ferskvandsfiskenes utbredelse og indvandring i Norge, med et tillæg om krebsen*. Centraltrykkeriet.

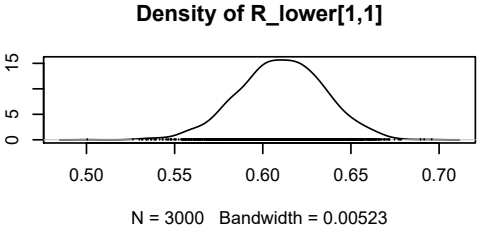
Pebesma, E. (2018). sf: Simple Features for R. *R Package Version 0. 6-0*.

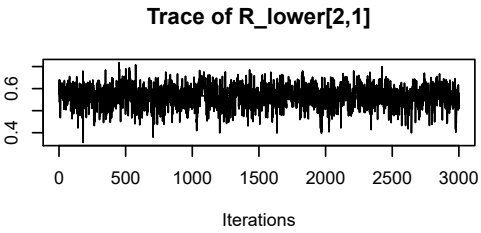
Sandlund, O. T., & Hesthagen, T. (2011). Fish diversity in Norwegian lakes: conserving species poor systems. In Jankun, M., Furghala-Selezniow, G., Wozniak, M., Wisniewska, AM. (Ed.), *Water Biodiversity Assessment and Protection* (pp. 7–20). University of Warmia and Mazury.

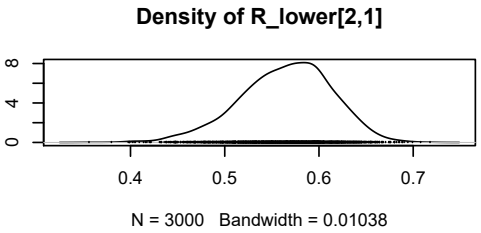
Supplementary Figures S3 for *Estimating community-level changes in freshwater species associations over a temperature gradient*

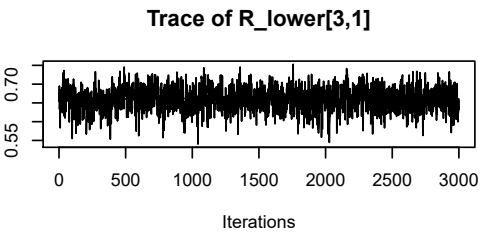
S3: Trace plots for species association parameters as defined in a large-scale Joint Species Distribution Model which accounted for changes in species association over a temperature gradient. Model based on species presence-absence data which included 15 species in 3308 lakes across the Fennoscandian region.

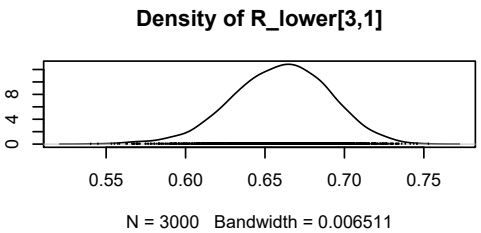

Trace of R_lower[1,1]
0.50
0.65
0
500
1000
1500
2000
2500
3000
Iterations

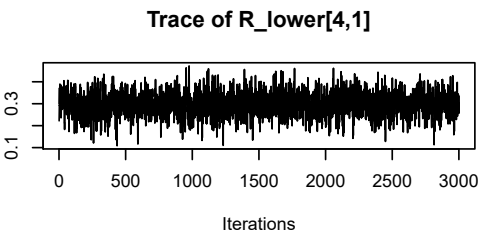

Density of R_lower[1,1]
0
5
15
0.50
0.55
0.60
0.65
0.70
N = 3000 Bandwidth = 0.00523

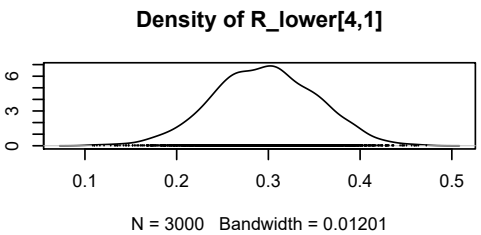

Trace of R_lower[2,1]
0.4
0.6
0
500
1000
1500
2000
2500
3000
Iterations

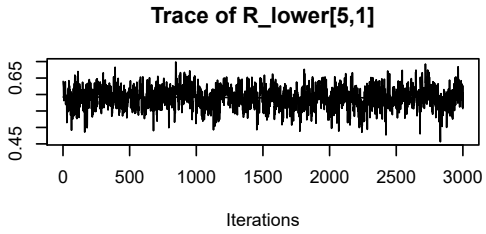

Density of R_lower[2,1]
0
4
8
0.4
0.5
0.6
0.7
N = 3000 Bandwidth = 0.01038

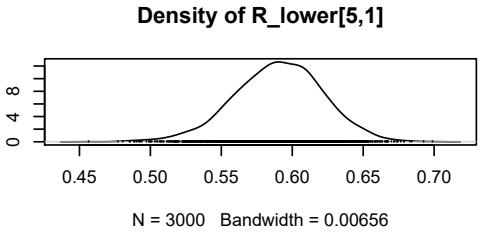

Trace of R_lower[3,1]
0.55
0.70
0
500
1000
1500
2000
2500
3000
Iterations

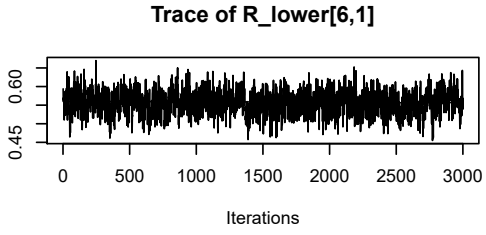

Density of R_lower[3,1]
0
4
8
0.55
0.60
0.65
0.70
0.75
N = 3000 Bandwidth = 0.006511

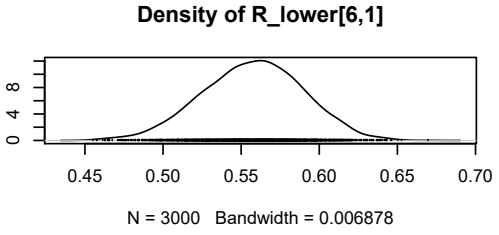

Trace of R_lower[4,1]
0.1
0.3
0
500
1000
1500
2000
2500
3000
Iterations

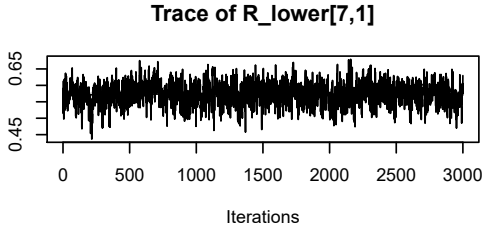

Density of R_lower[4,1]
0
3
6
0.1
0.2
0.3
0.4
0.5
N = 3000 Bandwidth = 0.01201

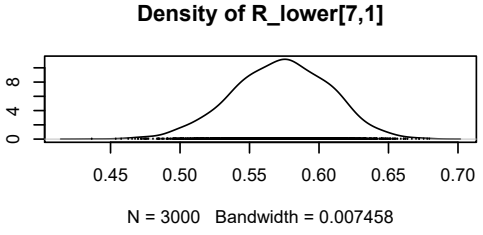

Trace of R_lower[5,1]
0.45
0.65
0
500
1000
1500
2000
2500
3000
Iterations

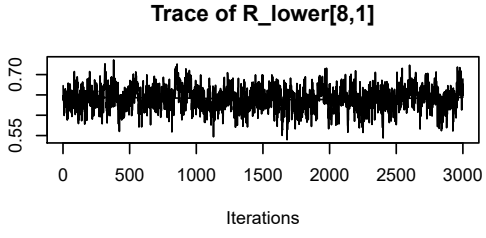

Density of R_lower[5,1]
0
4
8
0.45
0.50
0.55
0.60
0.65
0.70
N = 3000 Bandwidth = 0.00656

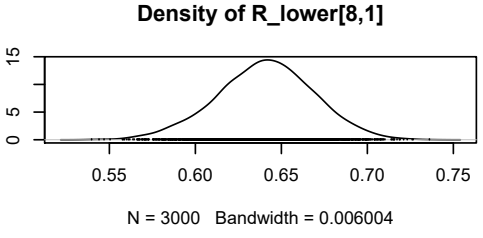

Trace of R_lower[6,1]
0.45
0.60
0
500
1000
1500
2000
2500
3000
Iterations

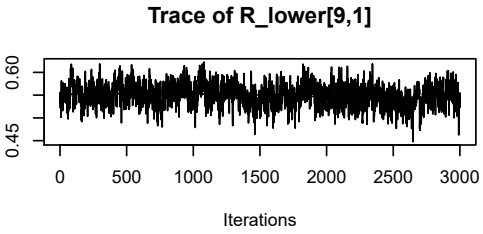

Density of R_lower[6,1]
0
4
8
0.45
0.50
0.55
0.60
0.65
0.70
N = 3000 Bandwidth = 0.006878

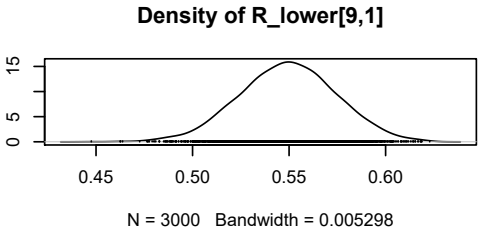

Trace of R_lower[7,1]
0.45
0.65
0
500
1000
1500
2000
2500
3000
Iterations

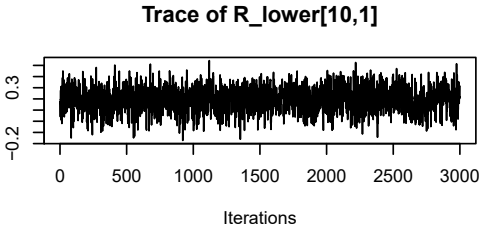

Density of R_lower[7,1]
0
4
8
0.45
0.50
0.55
0.60
0.65
0.70
N = 3000 Bandwidth = 0.007458

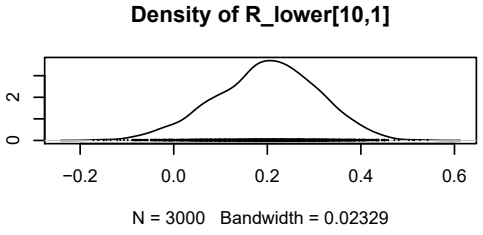

Trace of R_lower[8,1]
0.55
0.70
0
500
1000
1500
2000
2500
3000
Iterations

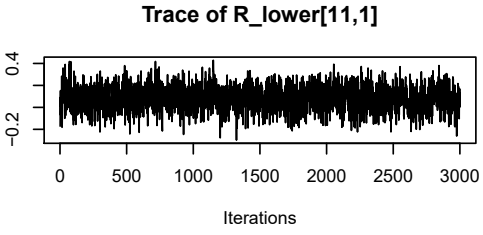

Density of R_lower[8,1]
0
5
15
0.55
0.60
0.65
0.70
0.75
N = 3000 Bandwidth = 0.006004

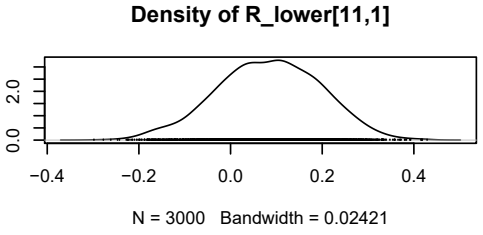

Trace of R_lower[9,1]
0.45
0.60
0
500
1000
1500
2000
2500
3000
Iterations

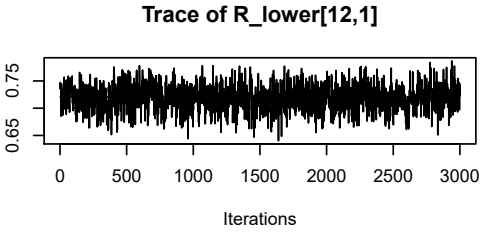

Density of R_lower[9,1]
0
5
15
0.45
0.50
0.55
0.60
N = 3000 Bandwidth = 0.005298

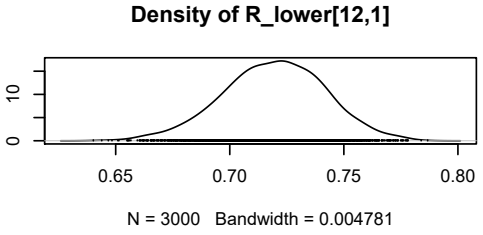

Trace of R_lower[10,1]
−0.2
0.3
0
500
1000
1500
2000
2500
3000
Iterations

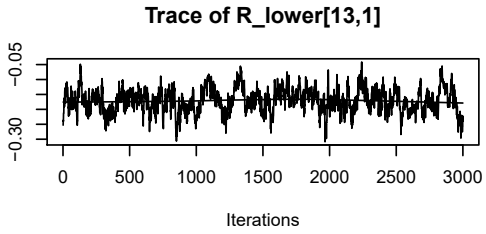

Density of R_lower[10,1]
0
2
−0.2
0.0
0.2
0.4
0.6
N = 3000 Bandwidth = 0.02329

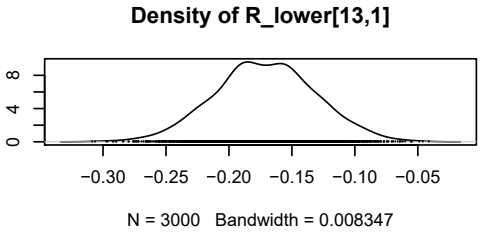

Trace of R_lower[11,1]
−0.2
0.4
0
500
1000
1500
2000
2500
3000
Iterations

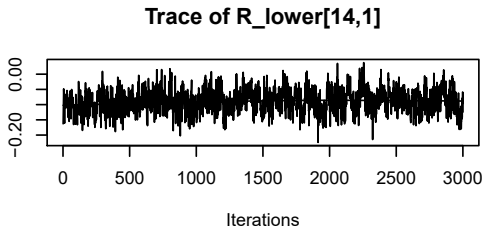

Density of R_lower[11,1]
0.0
2.0
−0.4
−0.2
0.0
0.2
0.4
N = 3000 Bandwidth = 0.02421

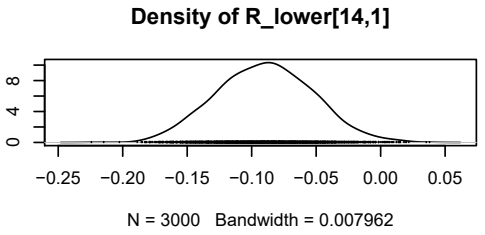

Trace of R_lower[12,1]
0.65
0.75
0
500
1000
1500
2000
2500
3000
Iterations

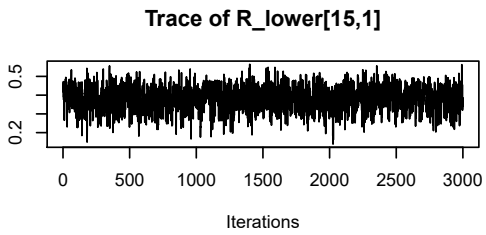

Density of R_lower[12,1]
0
10
0.65
0.70
0.75
0.80
N = 3000 Bandwidth = 0.004781

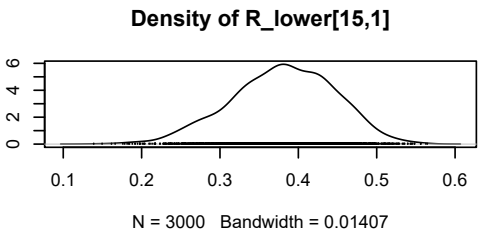

Trace of R_lower[13,1]
−0.05
−0.30
0
500
1000
1500
2000
2500
3000
Iterations

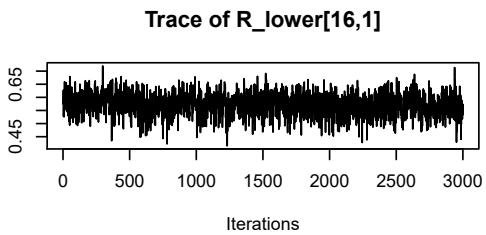

Density of R_lower[13,1]
0
4
8
−0.30
−0.25
−0.20
−0.15
−0.10
−0.05
N = 3000 Bandwidth = 0.008347

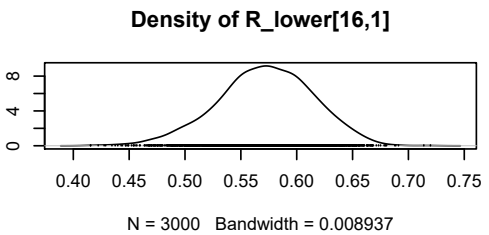

Trace of R_lower[14,1]
0.00
−0.20
0
500
1000
1500
2000
2500
3000
Iterations

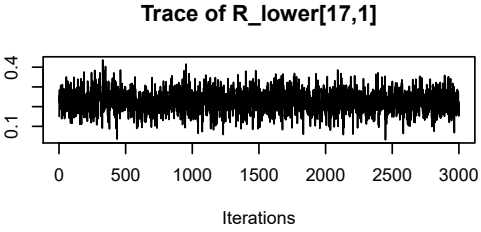

Density of R_lower[14,1]
0
4
8
−0.25
−0.20
−0.15
−0.10
−0.05
0.00
0.05
N = 3000 Bandwidth = 0.007962

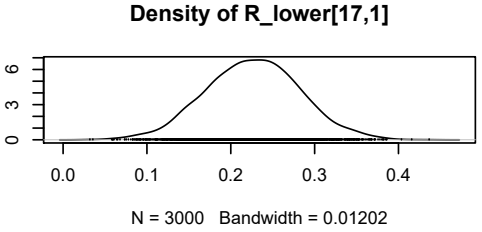

Trace of R_lower[15,1]
0.5
0.2
0
500
1000
1500
2000
2500
3000
Iterations

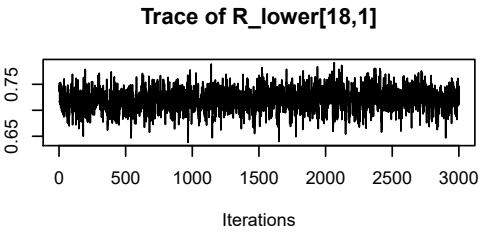

Density of R_lower[15,1]
0
2
4
6
0.1
0.2
0.3
0.4
0.5
0.6
N = 3000 Bandwidth = 0.01407

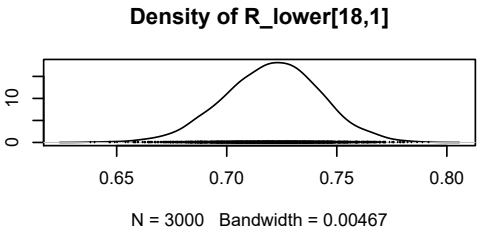

Trace of R_lower[16,1]
0.65
0.45
0
500
1000
1500
2000
2500
3000
Iterations

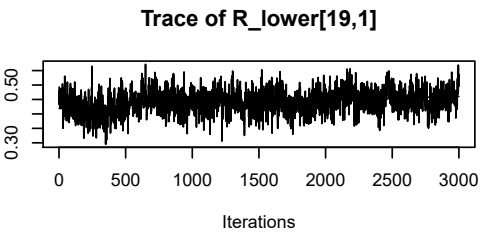

Density of R_lower[16,1]
0
4
8
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
N = 3000 Bandwidth = 0.008937

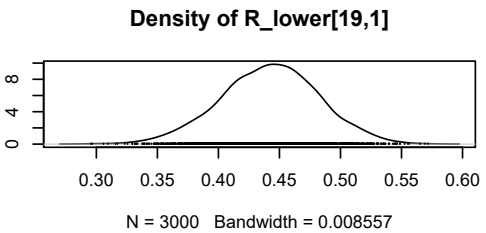

Trace of R_lower[17,1]
0.1
0.4
0
500
1000
1500
2000
2500
3000
Iterations

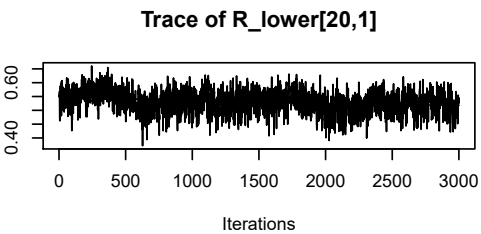

Density of R_lower[17,1]
0
3
6
0.0
0.1
0.2
0.3
0.4
N = 3000 Bandwidth = 0.01202

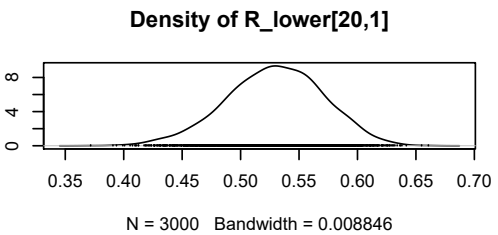

Trace of R_lower[18,1]
0.65
0.75
0
500
1000
1500
2000
2500
3000
Iterations

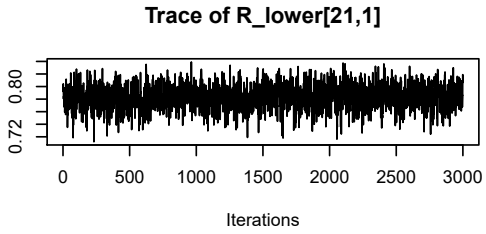

Density of R_lower[18,1]
0
10
0.65
0.70
0.75
0.80
N = 3000 Bandwidth = 0.00467

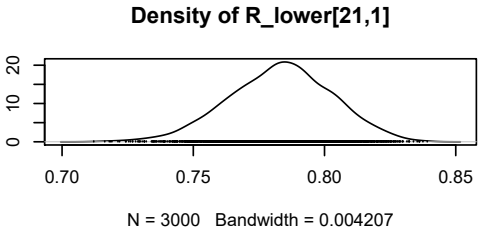

Trace of R_lower[19,1]
0.30
0.50
0
500
1000
1500
2000
2500
3000
Iterations

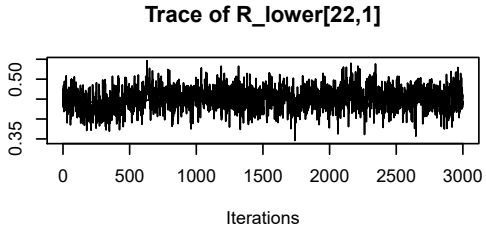

Density of R_lower[19,1]
0
4
8
0.30
0.35
0.40
0.45
0.50
0.55
0.60
N = 3000 Bandwidth = 0.008557

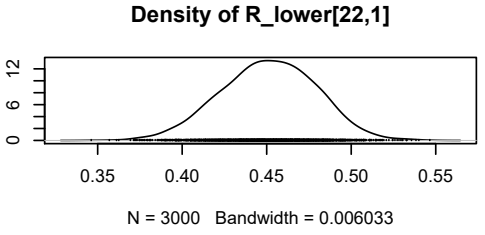

Trace of R_lower[20,1]
0.40
0.60
0
500
1000
1500
2000
2500
3000
Iterations

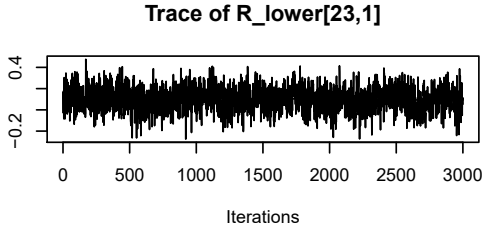

Density of R_lower[20,1]
0
4
8
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
N = 3000 Bandwidth = 0.008846

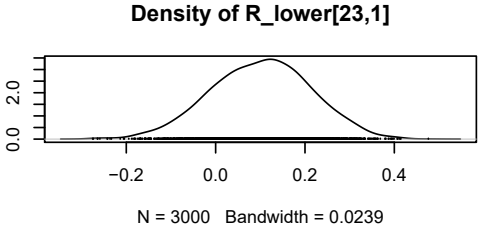

Trace of R_lower[21,1]
0.72
0.80
0
500
1000
1500
2000
2500
3000
Iterations

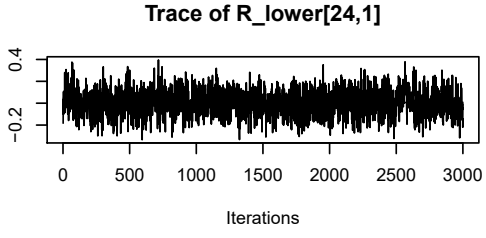

Density of R_lower[21,1]
0
10
20
0.70
0.75
0.80
0.85
N = 3000 Bandwidth = 0.004207

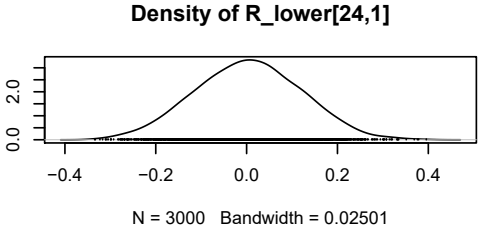

Trace of R_lower[22,1]
0.35
0.50
0
500
1000
1500
2000
2500
3000
Iterations

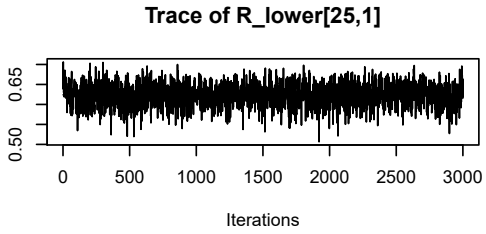

Density of R_lower[22,1]
0
6
12
0.35
0.40
0.45
0.50
0.55
N = 3000 Bandwidth = 0.006033

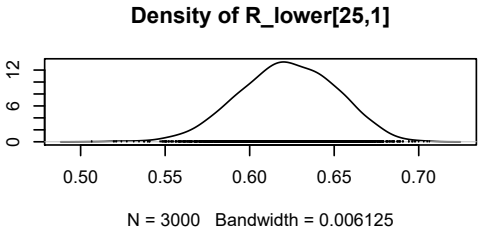

Trace of R_lower[23,1]
−0.2
0.4
0
500
1000
1500
2000
2500
3000
Iterations

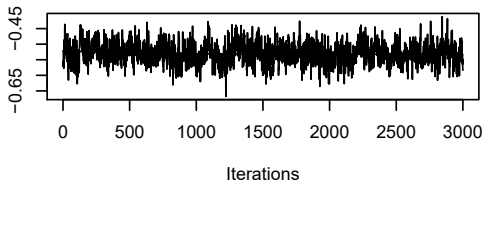

Density of R_lower[23,1]
0.0
2.0
−0.2
0.0
0.2
0.4
N = 3000 Bandwidth = 0.0239

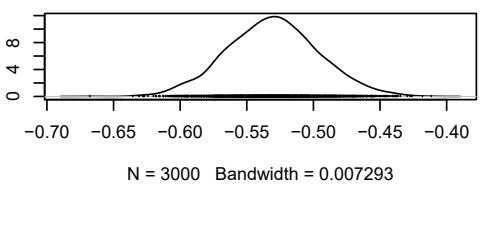

Trace of R_lower[24,1]
−0.2
0.4
0
500
1000
1500
2000
2500
3000
Iterations

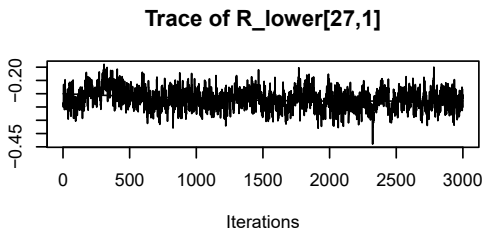

Density of R_lower[24,1]
0.0
2.0
−0.4
−0.2
0.0
0.2
0.4
N = 3000 Bandwidth = 0.02501

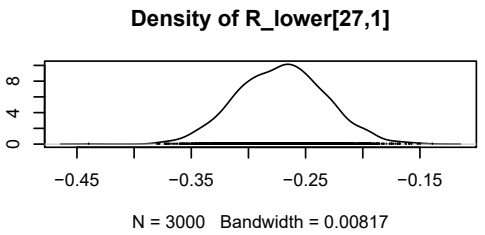

Trace of R_lower[25,1]
0.50
0.65
0
500
1000
1500
2000
2500
3000
Iterations

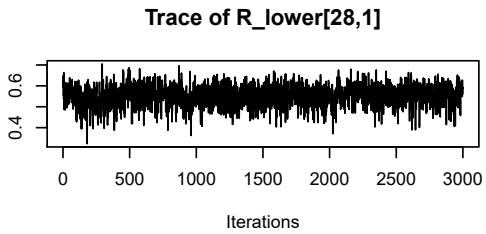

Density of R_lower[25,1]
0
6
12
0.50
0.55
0.60
0.65
0.70
N = 3000 Bandwidth = 0.006125

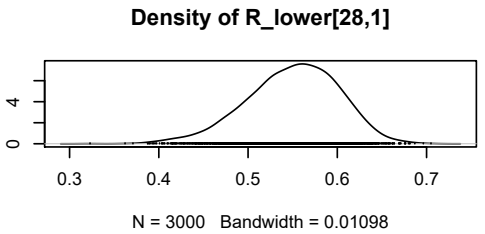

Trace of R_lower[26,1]
−0.65
−0.45
0
500
1000
1500
2000
2500
3000
Iterations

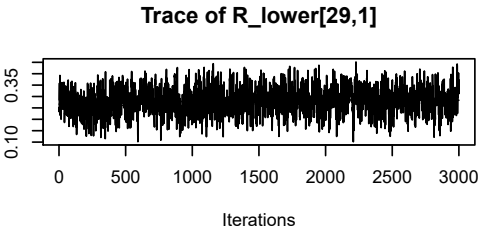

Density of R_lower[26,1]
0
4
8
−0.70
−0.65
−0.60
−0.55
−0.50
−0.45
−0.40
N = 3000 Bandwidth = 0.007293

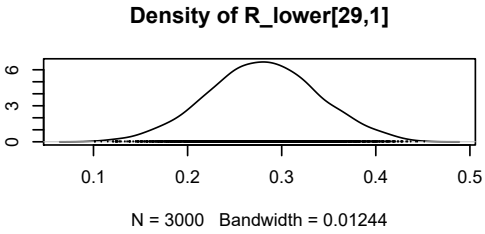

Trace of R_lower[27,1]
−0.45
−0.20
0
500
1000
1500
2000
2500
3000
Iterations

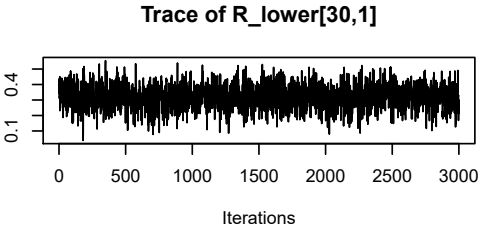

Density of R_lower[27,1]
0
4
8
−0.45
−0.35
−0.25
−0.15
N = 3000 Bandwidth = 0.00817

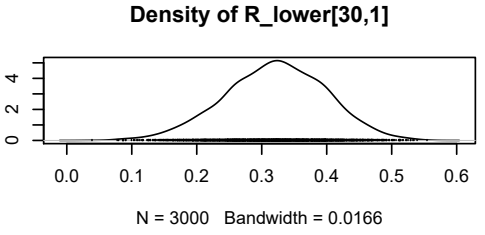

Trace of R_lower[28,1]
0.4
0.6
0
500
1000
1500
2000
2500
3000
Iterations

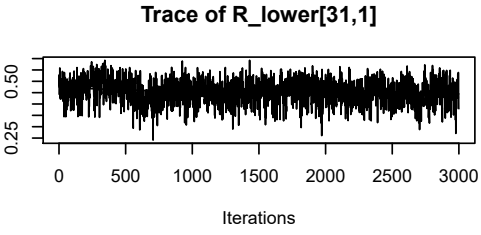

Density of R_lower[28,1]
0
4
0.3
0.4
0.5
0.6
0.7
N = 3000 Bandwidth = 0.01098

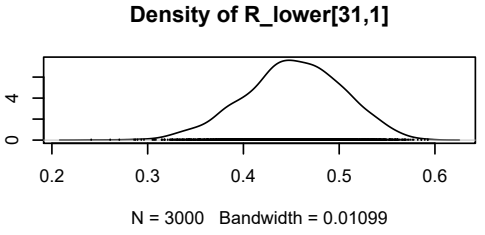

Trace of R_lower[29,1]
0.10
0.35
0
500
1000
1500
2000
2500
3000
Iterations

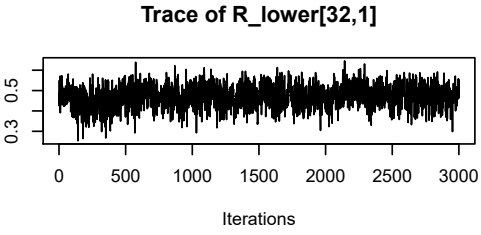

Density of R_lower[29,1]
0
3
6
0.1
0.2
0.3
0.4
0.5
N = 3000 Bandwidth = 0.01244

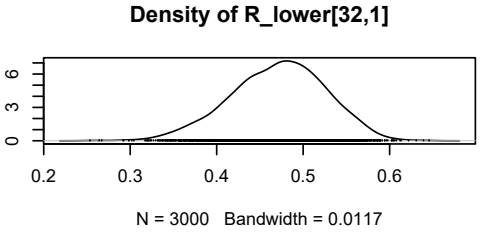

Trace of R_lower[30,1]
0.1
0.4
0
500
1000
1500
2000
2500
3000
Iterations

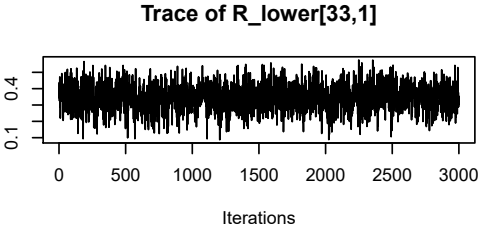

Density of R_lower[30,1]
0
2
4
0.0
0.1
0.2
0.3
0.4
0.5
0.6
N = 3000 Bandwidth = 0.0166

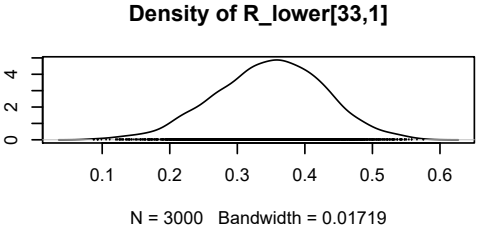

Trace of R_lower[31,1]
0.25
0.50
0
500
1000
1500
2000
2500
3000
Iterations

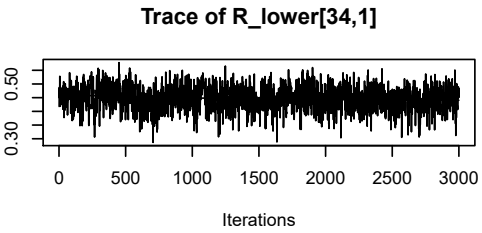

Density of R_lower[31,1]
0
4
0.2
0.3
0.4
0.5
0.6
N = 3000 Bandwidth = 0.01099

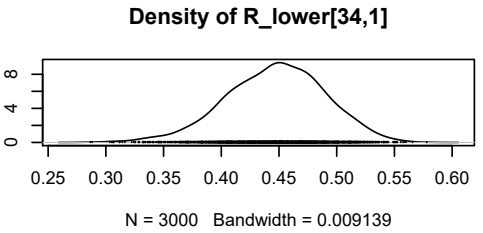

Trace of R_lower[32,1]
0.3
0.5
0
500
1000
1500
2000
2500
3000
Iterations

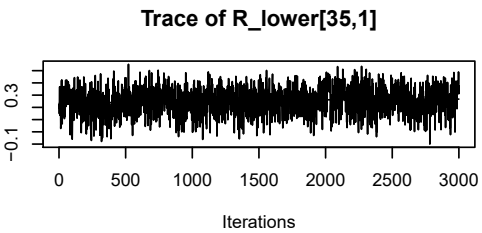

Density of R_lower[32,1]
0
3
6
0.2
0.3
0.4
0.5
0.6
N = 3000 Bandwidth = 0.0117

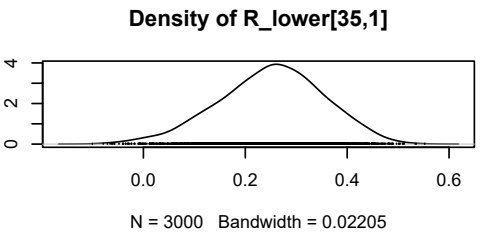

Trace of R_lower[33,1]
0.1
0.4
0
500
1000
1500
2000
2500
3000
Iterations

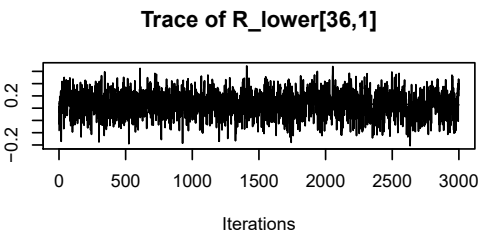

Density of R_lower[33,1]
0
2
4
0.1
0.2
0.3
0.4
0.5
0.6
N = 3000 Bandwidth = 0.01719

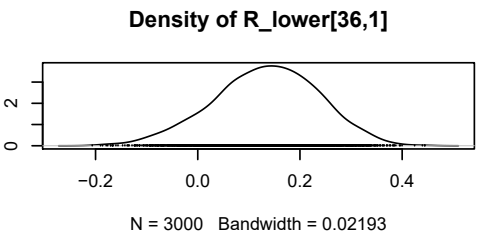

Trace of R_lower[34,1]
0.30
0.50
0
500
1000
1500
2000
2500
3000
Iterations

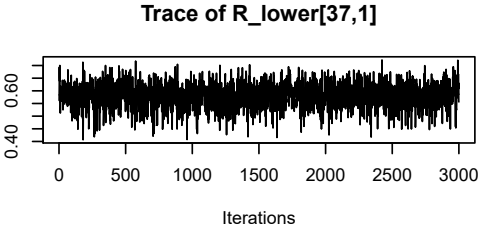

Density of R_lower[34,1]
0
4
8
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
N = 3000 Bandwidth = 0.009139

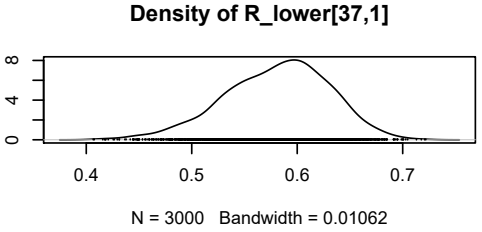

Trace of R_lower[35,1]
−0.1
0.3
0
500
1000
1500
2000
2500
3000
Iterations

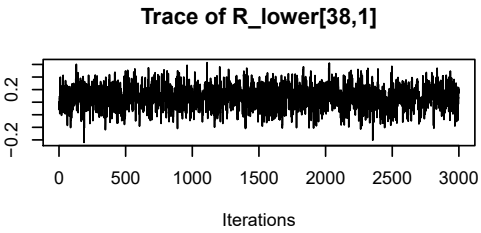

Density of R_lower[35,1]
0
2
4
0.0
0.2
0.4
0.6
N = 3000 Bandwidth = 0.02205

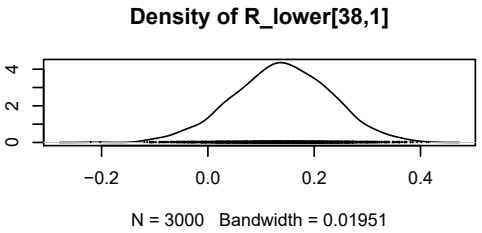

Trace of R_lower[36,1]
−0.2
0.2
0
500
1000
1500
2000
2500
3000
Iterations

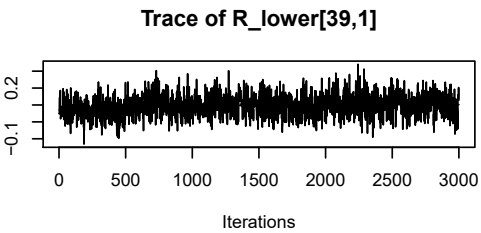

Density of R_lower[36,1]
0
2
−0.2
0.0
0.2
0.4
N = 3000 Bandwidth = 0.02193

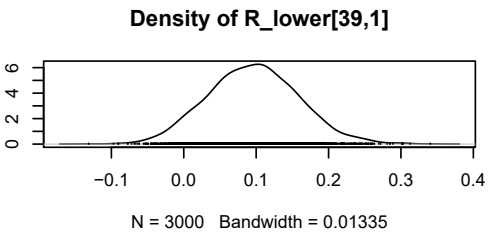

Trace of R_lower[37,1]
0.40
0.60
0
500
1000
1500
2000
2500
3000
Iterations

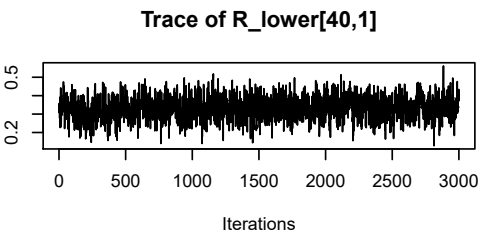

Density of R_lower[37,1]
0
4
8
0.4
0.5
0.6
0.7
N = 3000 Bandwidth = 0.01062

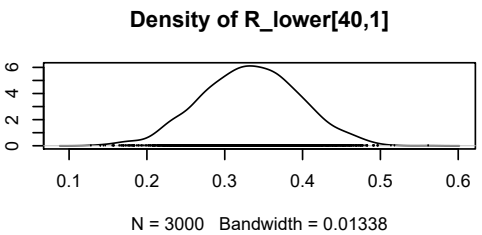

Trace of R_lower[38,1]
−0.2
0.2
0
500
1000
1500
2000
2500
3000
Iterations

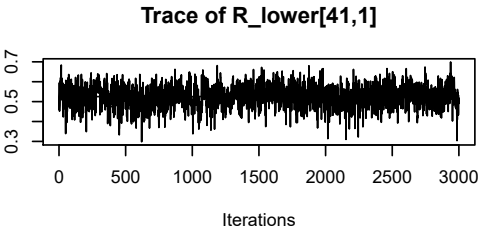

Density of R_lower[38,1]
0
2
4
−0.2
0.0
0.2
0.4
N = 3000 Bandwidth = 0.01951

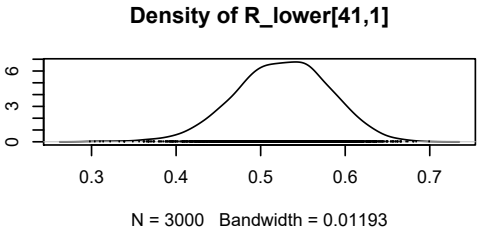

Trace of R_lower[39,1]
−0.1
0.2
0
500
1000
1500
2000
2500
3000
Iterations

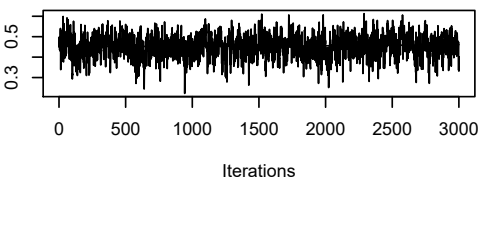

Density of R_lower[39,1]
0
2
4
6
−0.1
0.0
0.1
0.2
0.3
0.4
N = 3000 Bandwidth = 0.01335

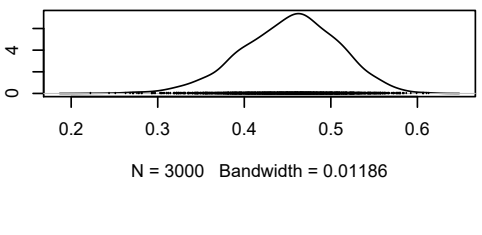

Trace of R_lower[40,1]
0.2
0.5
0
500
1000
1500
2000
2500
3000
Iterations

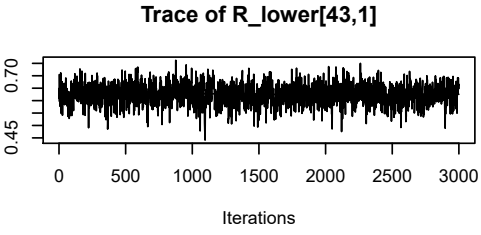

Density of R_lower[40,1]
0
2
4
6
0.1
0.2
0.3
0.4
0.5
0.6
N = 3000 Bandwidth = 0.01338

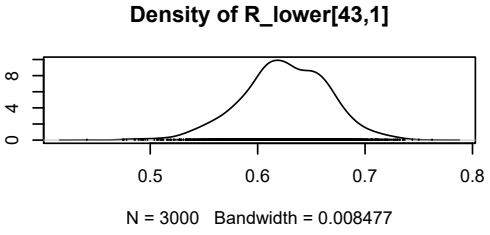

Trace of R_lower[41,1]
0.3 0.5 0.7
0 500 1000 1500 2000 2500 3000
Iterations

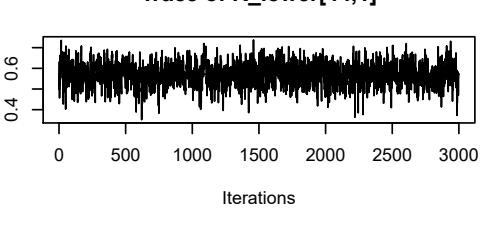

Density of R_lower[41,1]
0 3 6
0.3 0.4 0.5 0.6 0.7
N = 3000 Bandwidth = 0.01193

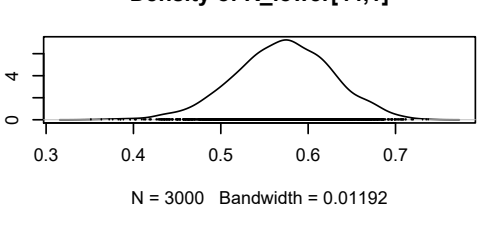

Trace of R_lower[42,1]
0.3 0.5
0 500 1000 1500 2000 2500 3000
Iterations

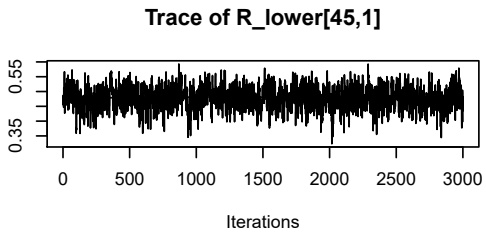

Density of R_lower[42,1]
0 4
0.2 0.3 0.4 0.5 0.6
N = 3000 Bandwidth = 0.01186

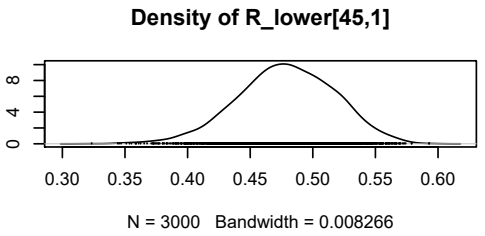

Trace of R_lower[43,1]
0.45 0.70
0 500 1000 1500 2000 2500 3000
Iterations

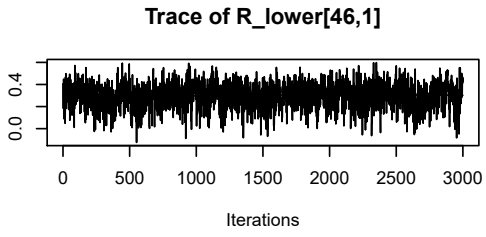

Density of R_lower[43,1]
0 4 8
0.5 0.6 0.7 0.8
N = 3000 Bandwidth = 0.008477

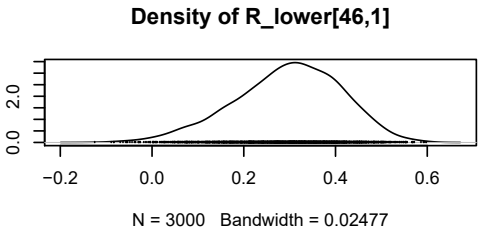

Trace of R_lower[44,1]
0.4 0.6
0 500 1000 1500 2000 2500 3000
Iterations

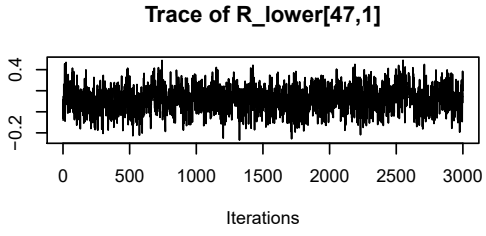

Density of R_lower[44,1]
0 4
0.3 0.4 0.5 0.6 0.7
N = 3000 Bandwidth = 0.01192

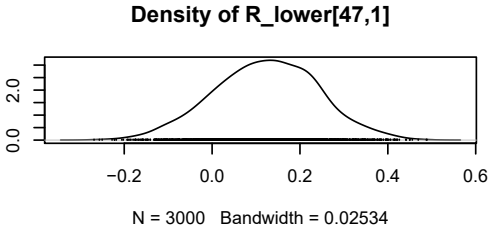

Trace of R_lower[45,1]
0.35
0.55
0
500
1000
1500
2000
2500
3000
Iterations

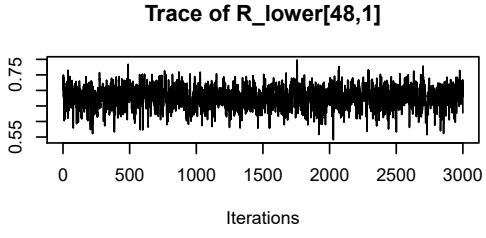

Density of R_lower[45,1]
0
4
8
0.30
0.35
0.40
0.45
0.50
0.55
0.60
N = 3000 Bandwidth = 0.008266

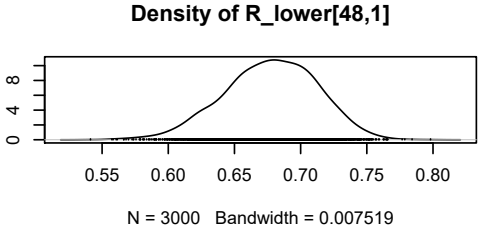

Trace of R_lower[46,1]
0.0
0.4
0
500
1000
1500
2000
2500
3000
Iterations

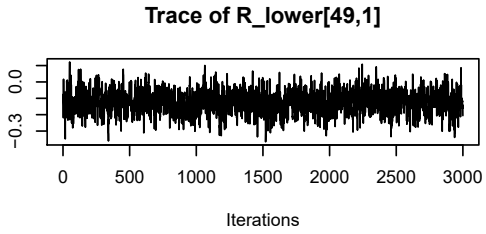

Density of R_lower[46,1]
0.0
2.0
−0.2
0.0
0.2
0.4
0.6
N = 3000 Bandwidth = 0.02477

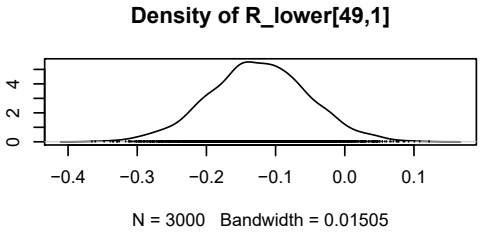

Trace of R_lower[47,1]
−0.2
0.4
0
500
1000
1500
2000
2500
3000
Iterations

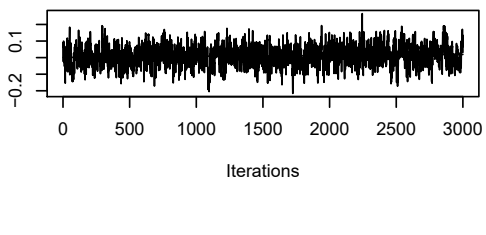

Density of R_lower[47,1]
0.0
2.0
−0.2
0.0
0.2
0.4
0.6
N = 3000 Bandwidth = 0.02534

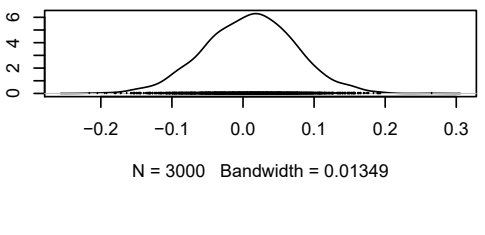

Trace of R_lower[48,1]
0.55
0.75
0
500
1000
1500
2000
2500
3000
Iterations

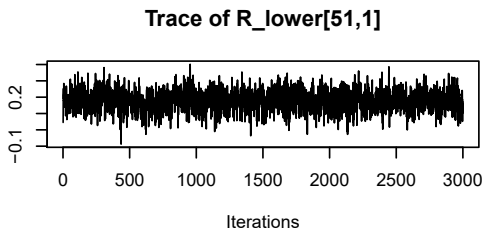

Density of R_lower[48,1]
0
4
8
0.55
0.60
0.65
0.70
0.75
0.80
N = 3000 Bandwidth = 0.007519

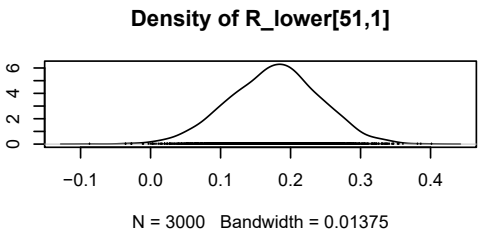

Trace of R_lower[49,1]
−0.3 0.0
0 500 1000 1500 2000 2500 3000
Iterations

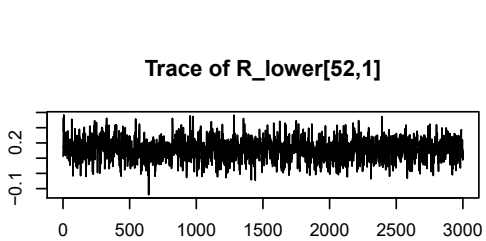

Density of R_lower[49,1]
0 2 4
−0.4 −0.3 −0.2 −0.1 0.0 0.1
N = 3000 Bandwidth = 0.01505

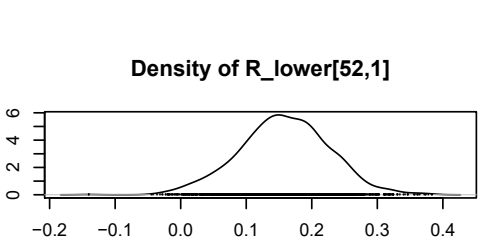

Trace of R_lower[50,1]
−0.2 0.1
0 500 1000 1500 2000 2500 3000
Iterations

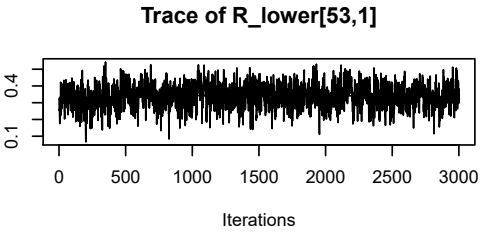

Density of R_lower[50,1]
0 2 4 6
−0.2 −0.1 0.0 0.1 0.2 0.3
N = 3000 Bandwidth = 0.01349

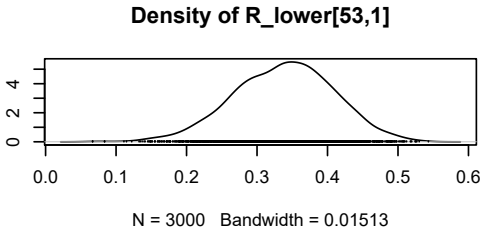

Trace of R_lower[51,1]
−0.1 0.2
0 500 1000 1500 2000 2500 3000
Iterations

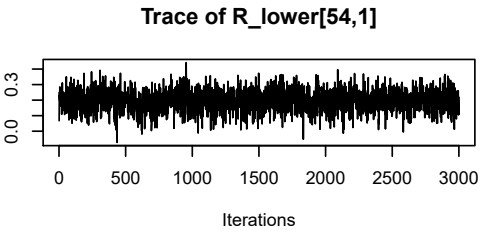

Density of R_lower[51,1]
0 2 4 6
−0.1 0.0 0.1 0.2 0.3 0.4
N = 3000 Bandwidth = 0.01375

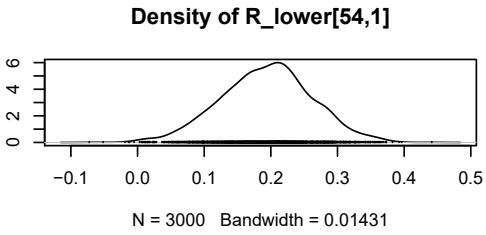

Trace of R_lower[52,1]
−0.1 0.2
0 500 1000 1500 2000 2500 3000
Iterations

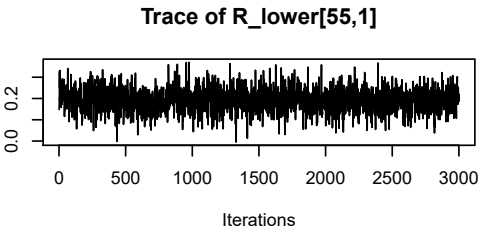

Density of R_lower[52,1]
0 2 4 6
−0.2 −0.1 0.0 0.1 0.2 0.3 0.4
N = 3000 Bandwidth = 0.0143

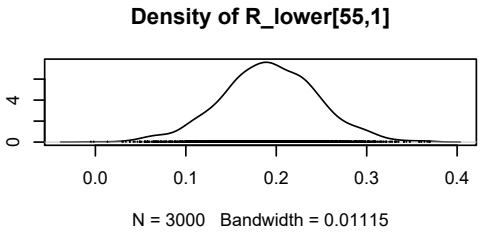

Trace of R_lower[53,1]
0.1
0.4
0
500
1000
1500
2000
2500
3000
Iterations

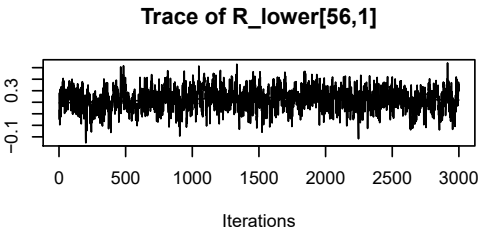

Density of R_lower[53,1]
0
2
4
0.0
0.1
0.2
0.3
0.4
0.5
0.6
N = 3000 Bandwidth = 0.01513

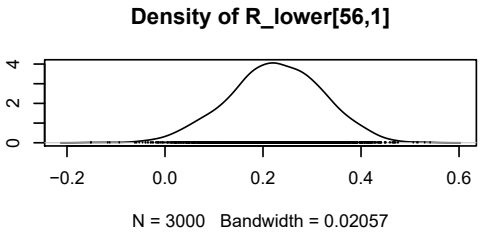

Trace of R_lower[54,1]
0.0
0.3
0
500
1000
1500
2000
2500
3000
Iterations

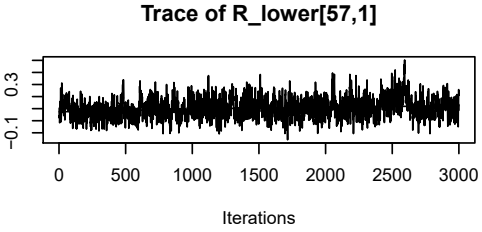

Density of R_lower[54,1]
0
2
4
6
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
N = 3000 Bandwidth = 0.01431

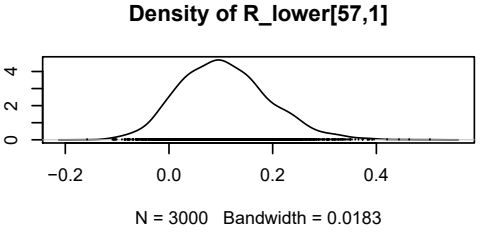

Trace of R_lower[55,1]
0.0
0.2
0
500
1000
1500
2000
2500
3000
Iterations

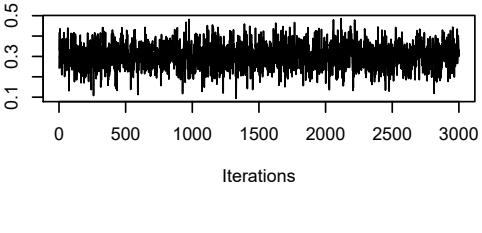

Density of R_lower[55,1]
0
4
0.0
0.1
0.2
0.3
0.4
N = 3000 Bandwidth = 0.01115

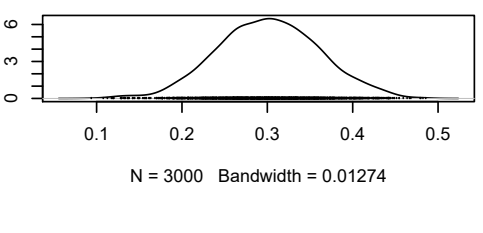

Trace of R_lower[56,1]
−0.1
0.3
0
500
1000
1500
2000
2500
3000
Iterations

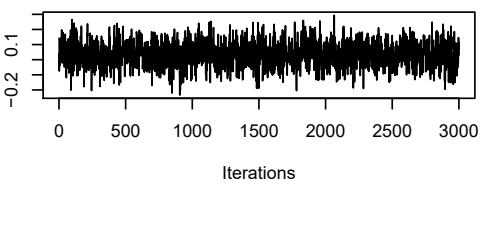

Density of R_lower[56,1]
0
2
4
−0.2
0.0
0.2
0.4
0.6
N = 3000 Bandwidth = 0.02057

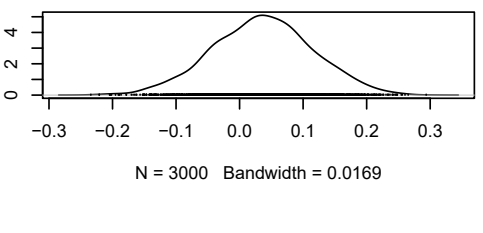

Trace of R_lower[57,1]
−0.1 0.3
0 500 1000 1500 2000 2500 3000
Iterations

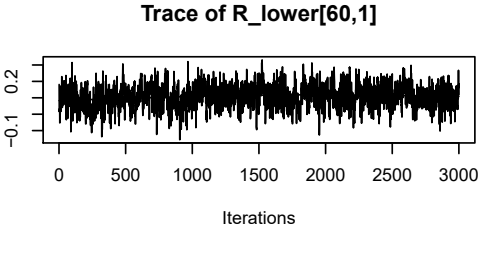

Density of R_lower[57,1]
0 2 4
−0.2 0.0 0.2 0.4
N = 3000 Bandwidth = 0.0183

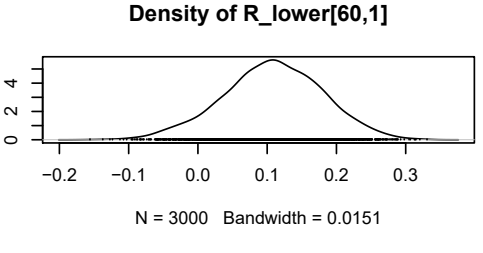

Trace of R_lower[58,1]
0.1 0.3 0.5
0 500 1000 1500 2000 2500 3000
Iterations

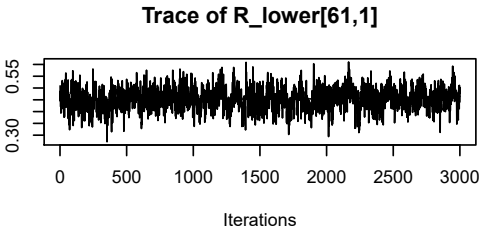

Density of R_lower[58,1]
0 3 6
0.1 0.2 0.3 0.4 0.5
N = 3000 Bandwidth = 0.01274

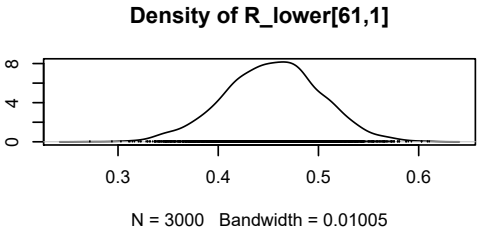

Trace of R_lower[59,1]
−0.2 0.1
0 500 1000 1500 2000 2500 3000
Iterations

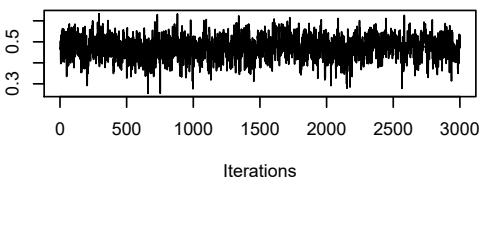

Density of R_lower[59,1]
0 2 4
−0.3 −0.2 −0.1 0.0 0.1 0.2 0.3
N = 3000 Bandwidth = 0.0169

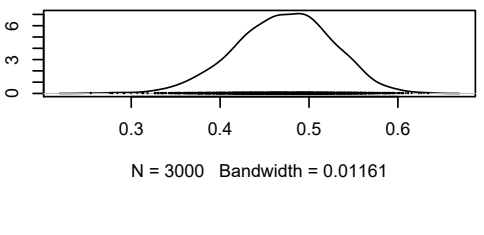

Trace of R_lower[60,1]
−0.1 0.2
0 500 1000 1500 2000 2500 3000
Iterations

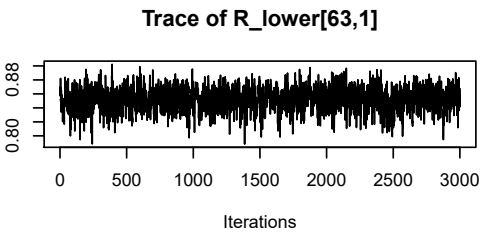

Density of R_lower[60,1]
0 2 4
−0.2 −0.1 0.0 0.1 0.2 0.3
N = 3000 Bandwidth = 0.0151

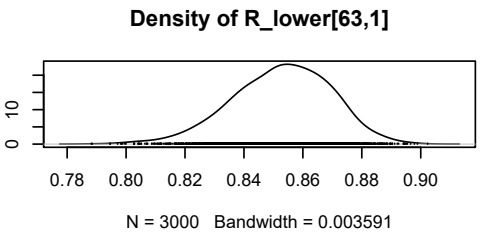

Trace of R_lower[61,1]
0.30
0.55
0
500
1000
1500
2000
2500
3000
Iterations

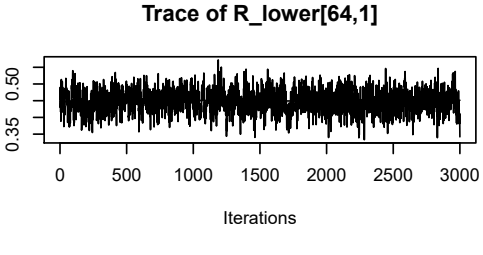

Density of R_lower[61,1]
0
4
8
0.3
0.4
0.5
0.6
N = 3000 Bandwidth = 0.01005

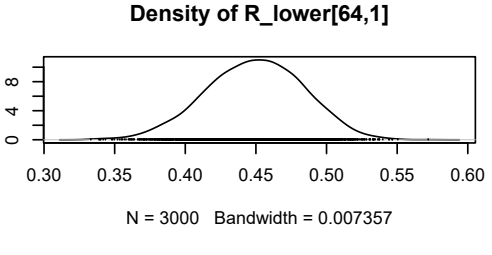

Trace of R_lower[62,1]
0.3
0.5
0
500
1000
1500
2000
2500
3000
Iterations

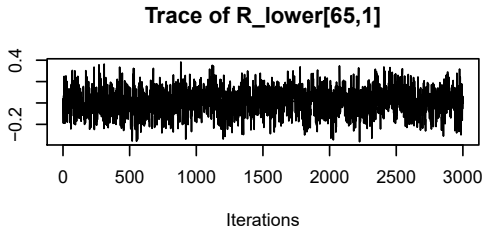

Density of R_lower[62,1]
0
3
6
0.3
0.4
0.5
0.6
N = 3000 Bandwidth = 0.01161

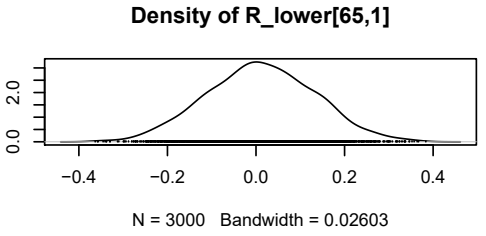

Trace of R_lower[63,1]
0.80
0.88
0
500
1000
1500
2000
2500
3000
Iterations

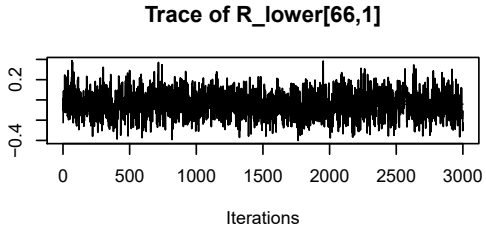

Density of R_lower[63,1]
0
10
0.78
0.80
0.82
0.84
0.86
0.88
0.90
N = 3000 Bandwidth = 0.003591

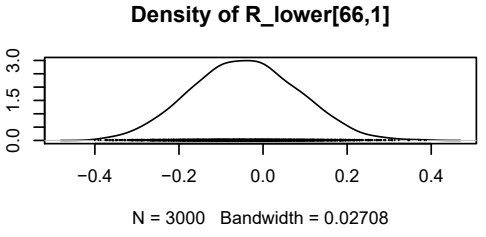

Trace of R_lower[64,1]
0.35
0.50
0
500
1000
1500
2000
2500
3000
Iterations

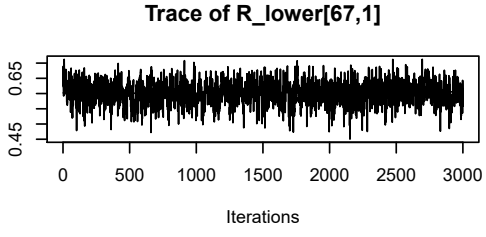

Density of R_lower[64,1]
0
4
8
0.30
0.35
0.40
0.45
0.50
0.55
0.60
N = 3000 Bandwidth = 0.007357

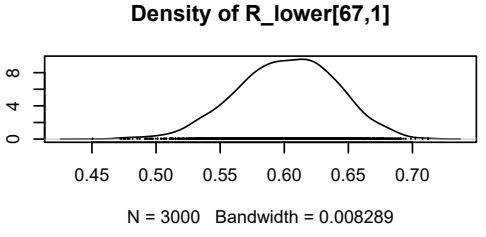

Trace of R_lower[65,1]
−0.2
0.4
0
500
1000
1500
2000
2500
3000
Iterations

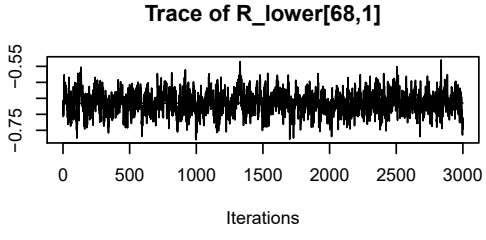

Density of R_lower[65,1]
0.0
2.0
−0.4
−0.2
0.0
0.2
0.4
N = 3000 Bandwidth = 0.02603

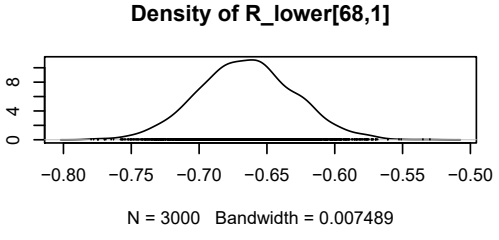

Trace of R_lower[66,1]
−0.4
0.2
0
500
1000
1500
2000
2500
3000
Iterations

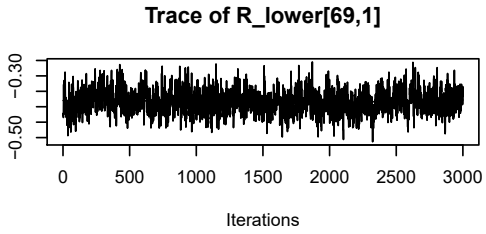

Density of R_lower[66,1]
0.0
1.5
3.0
−0.4
−0.2
0.0
0.2
0.4
N = 3000 Bandwidth = 0.02708

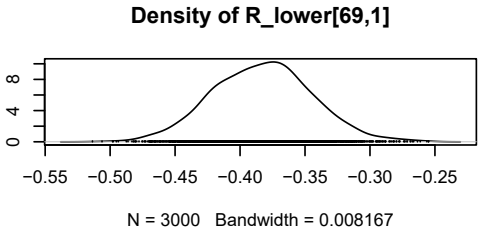

Trace of R_lower[67,1]
0.45
0.65
0
500
1000
1500
2000
2500
3000
Iterations

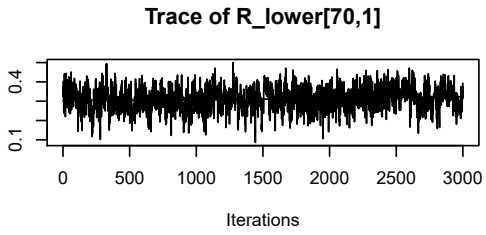

Density of R_lower[67,1]
0
4
8
0.45
0.50
0.55
0.60
0.65
0.70
N = 3000 Bandwidth = 0.008289

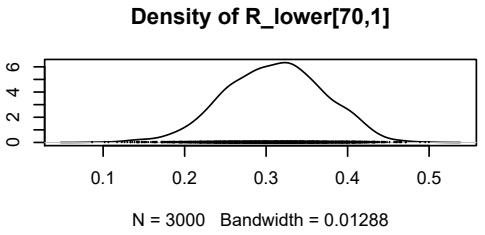

Trace of R_lower[68,1]
−0.75
−0.55
0
500
1000
1500
2000
2500
3000
Iterations

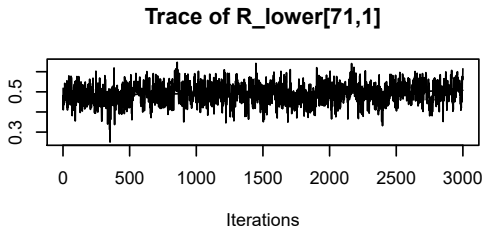

Density of R_lower[68,1]
0
4
8
−0.80
−0.75
−0.70
−0.65
−0.60
−0.55
−0.50
N = 3000 Bandwidth = 0.007489

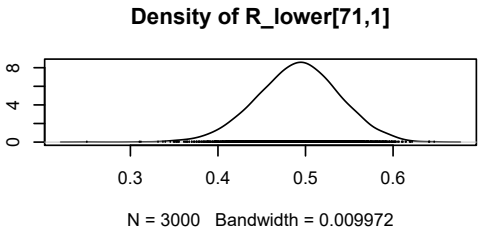

Trace of R_lower[69,1]
−0.50
−0.30
0
500
1000
1500
2000
2500
3000
Iterations

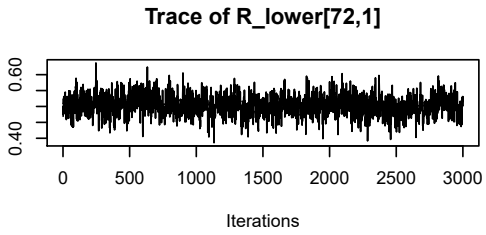

Density of R_lower[69,1]
0
4
8
−0.55
−0.50
−0.45
−0.40
−0.35
−0.30
−0.25
N = 3000 Bandwidth = 0.008167

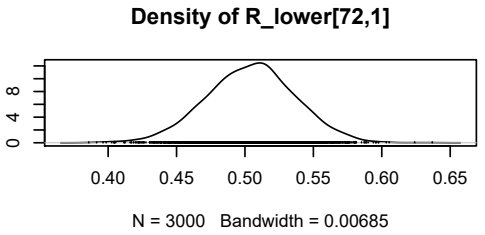

Trace of R_lower[70,1]
0.1
0.4
0
500
1000
1500
2000
2500
3000
Iterations

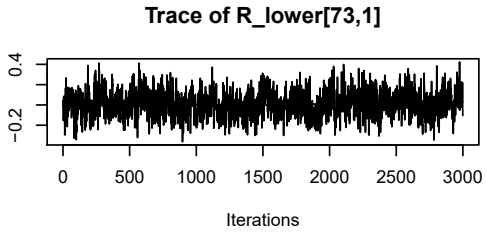

Density of R_lower[70,1]
0
2
4
6
0.1
0.2
0.3
0.4
0.5
N = 3000 Bandwidth = 0.01288

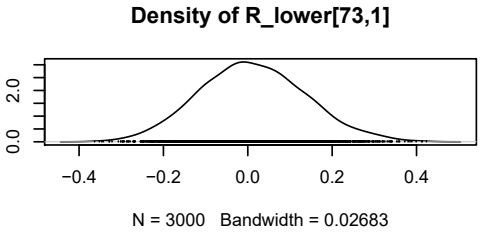

Trace of R_lower[71,1]
0.3
0.5
0
500
1000
1500
2000
2500
3000
Iterations

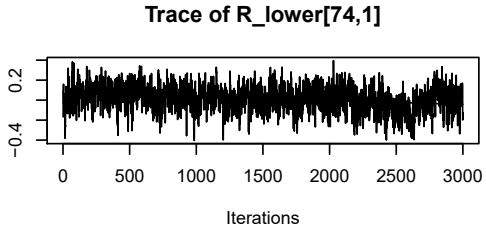

Density of R_lower[71,1]
0
4
8
0.3
0.4
0.5
0.6
N = 3000 Bandwidth = 0.009972

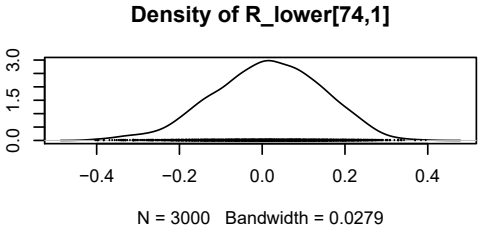

Trace of R_lower[72,1]
0.40
0.60
0
500
1000
1500
2000
2500
3000
Iterations

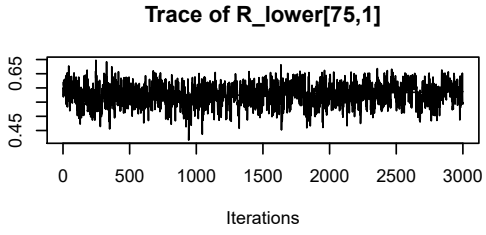

Density of R_lower[72,1]
0
4
8
0.40
0.45
0.50
0.55
0.60
0.65
N = 3000 Bandwidth = 0.00685

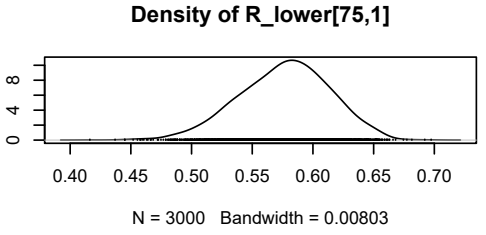

Trace of R_lower[73,1]
−0.2
0.4
0
500
1000
1500
2000
2500
3000
Iterations

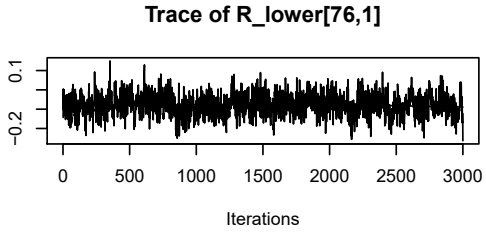

Density of R_lower[73,1]
0.0
2.0
−0.4
−0.2
0.0
0.2
0.4
N = 3000 Bandwidth = 0.02683

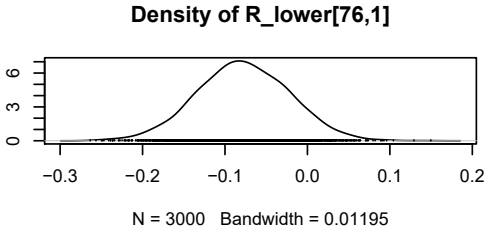

Trace of R_lower[74,1]
−0.4
0.2
0
500
1000
1500
2000
2500
3000
Iterations

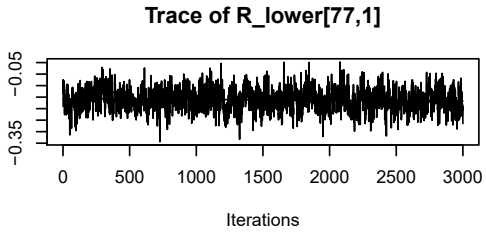

Density of R_lower[74,1]
0.0
1.5
3.0
−0.4
−0.2
0.0
0.2
0.4
N = 3000 Bandwidth = 0.0279

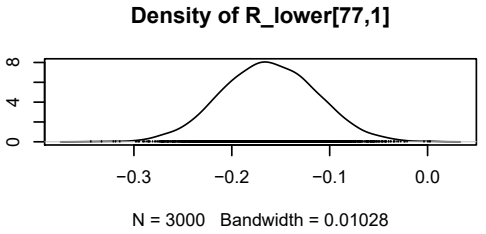

Trace of R_lower[75,1]
0.45
0.65
0
500
1000
1500
2000
2500
3000
Iterations

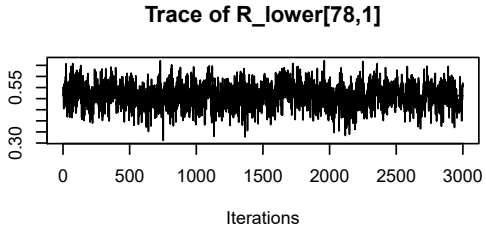

Density of R_lower[75,1]
0
4
8
0.40
0.45
0.50
0.55
0.60
0.65
0.70
N = 3000 Bandwidth = 0.00803

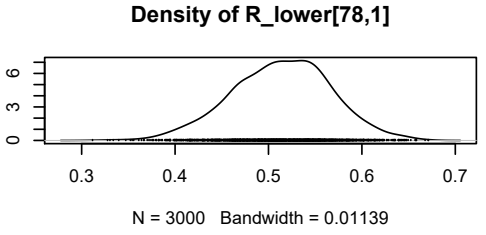

Trace of R_lower[76,1]
−0.2
0.1
0
500
1000
1500
2000
2500
3000
Iterations

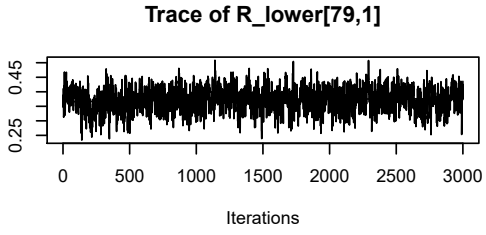

Density of R_lower[76,1]
0
3
6
−0.3
−0.2
−0.1
0.0
0.1
0.2
N = 3000 Bandwidth = 0.01195

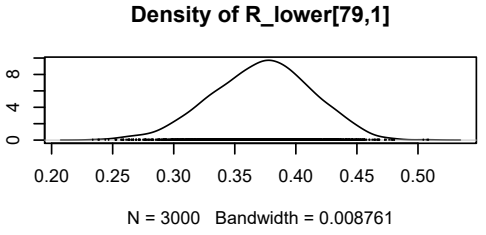

Trace of R_lower[77,1]
−0.05
−0.35
0
500
1000
1500
2000
2500
3000
Iterations

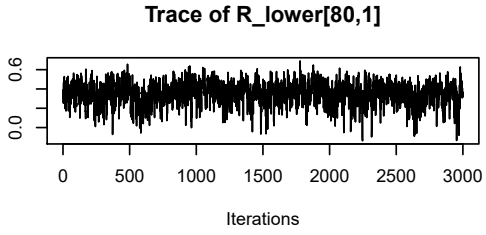

Density of R_lower[77,1]
8
4
0
−0.3
−0.2
−0.1
0.0
N = 3000 Bandwidth = 0.01028

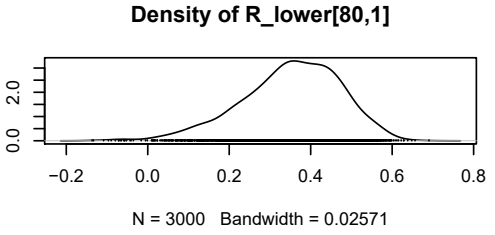

Trace of R_lower[78,1]
0.55
0.30
0
500
1000
1500
2000
2500
3000
Iterations

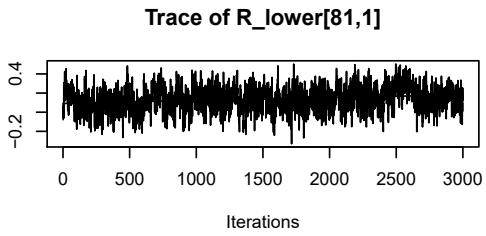

Density of R_lower[78,1]
6
3
0
0.3
0.4
0.5
0.6
0.7
N = 3000 Bandwidth = 0.01139

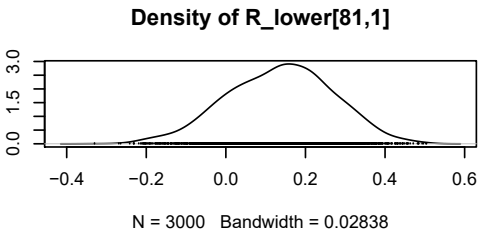

Trace of R_lower[79,1]
0.45
0.25
0
500
1000
1500
2000
2500
3000
Iterations

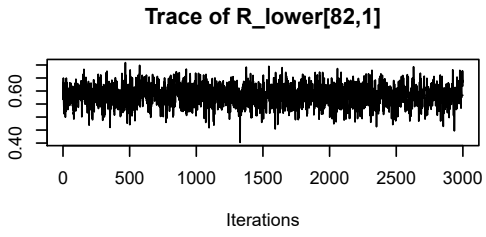

Density of R_lower[79,1]
8
4
0
0.20
0.25
0.30
0.35
0.40
0.45
0.50
N = 3000 Bandwidth = 0.008761

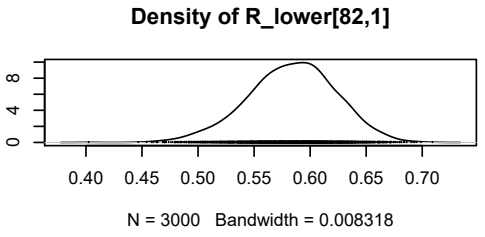

Trace of R_lower[80,1]
0.6
0.0
0
500
1000
1500
2000
2500
3000
Iterations

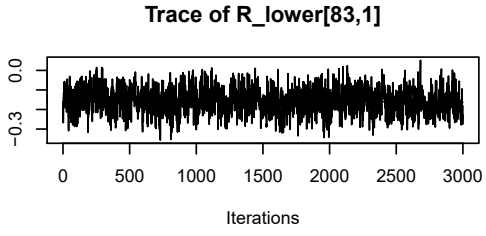

Density of R_lower[80,1]
2.0
0.0
−0.2
0.0
0.2
0.4
0.6
0.8
N = 3000 Bandwidth = 0.02571

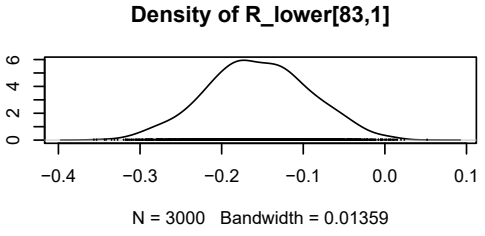

Trace of R_lower[81,1]
−0.2
0.4
0
500
1000
1500
2000
2500
3000
Iterations

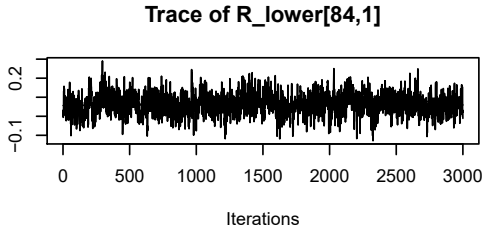

Density of R_lower[81,1]
0.0
1.5
3.0
−0.4
−0.2
0.0
0.2
0.4
0.6
N = 3000 Bandwidth = 0.02838

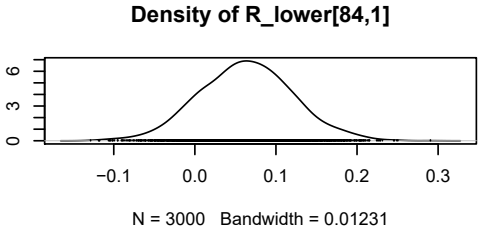

Trace of R_lower[82,1]
0.40
0.60
0
500
1000
1500
2000
2500
3000
Iterations

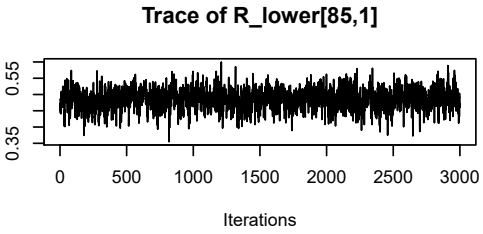

Density of R_lower[82,1]
0
4
8
0.40
0.45
0.50
0.55
0.60
0.65
0.70
N = 3000 Bandwidth = 0.008318

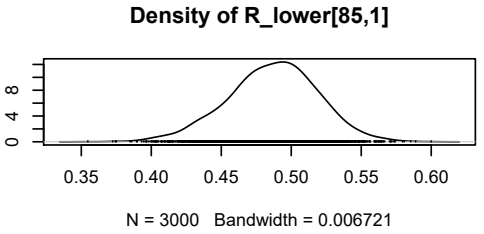

Trace of R_lower[83,1]
−0.3
0.0
0
500
1000
1500
2000
2500
3000
Iterations

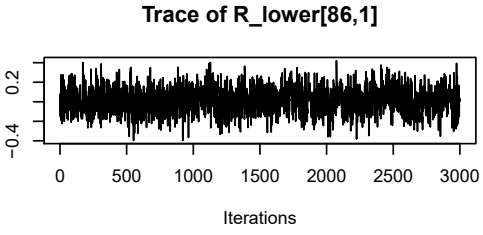

Density of R_lower[83,1]
0
2
4
6
−0.4
−0.3
−0.2
−0.1
0.0
0.1
N = 3000 Bandwidth = 0.01359

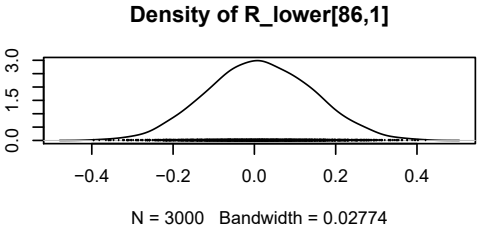

Trace of R_lower[84,1]
−0.1
0.2
0
500
1000
1500
2000
2500
3000
Iterations

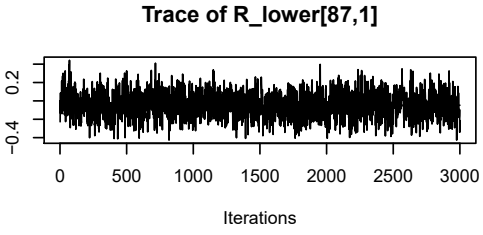

Density of R_lower[84,1]
0
3
6
−0.1
0.0
0.1
0.2
0.3
N = 3000 Bandwidth = 0.01231

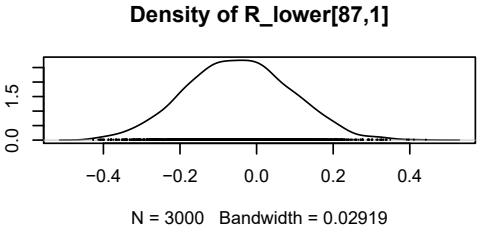

Trace of R_lower[85,1]
0.35
0.55
0
500
1000
1500
2000
2500
3000
Iterations

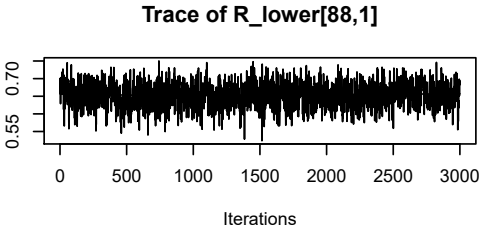

Density of R_lower[85,1]
0
4
8
0.35
0.40
0.45
0.50
0.55
0.60
N = 3000 Bandwidth = 0.006721

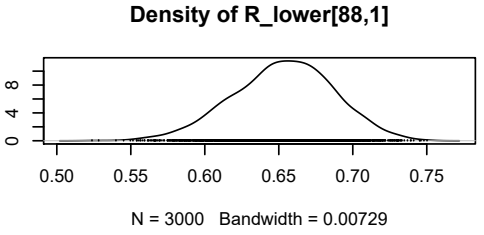

Trace of R_lower[86,1]
−0.4
0.2
0
500
1000
1500
2000
2500
3000
Iterations

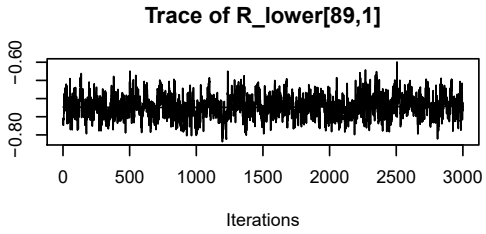

Density of R_lower[86,1]
0.0
1.5
3.0
−0.4
−0.2
0.0
0.2
0.4
N = 3000 Bandwidth = 0.02774

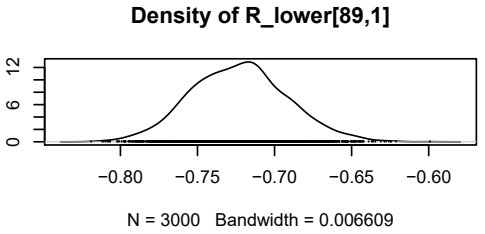

Trace of R_lower[87,1]
−0.4
0.2
0
500
1000
1500
2000
2500
3000
Iterations

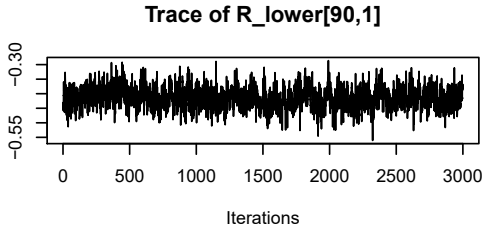

Density of R_lower[87,1]
0.0
1.5
−0.4
−0.2
0.0
0.2
0.4
N = 3000 Bandwidth = 0.02919

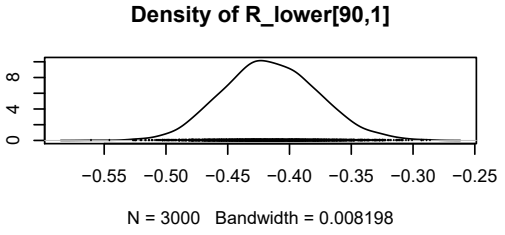

Trace of R_lower[88,1]
0.55
0.70
0
500
1000
1500
2000
2500
3000
Iterations

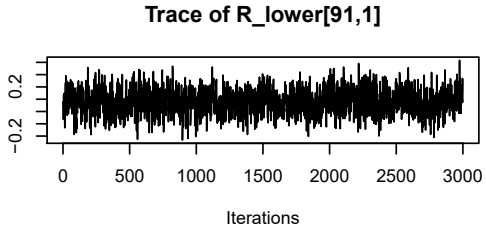

Density of R_lower[88,1]
0
4
8
0.50
0.55
0.60
0.65
0.70
0.75
N = 3000 Bandwidth = 0.00729

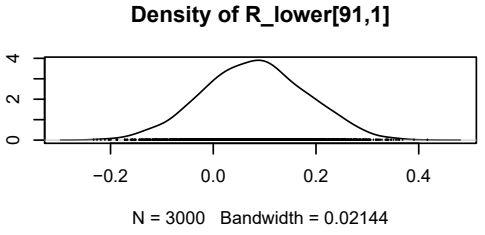

Trace of R_lower[89,1]
−0.60
−0.80
0
500
1000
1500
2000
2500
3000
Iterations

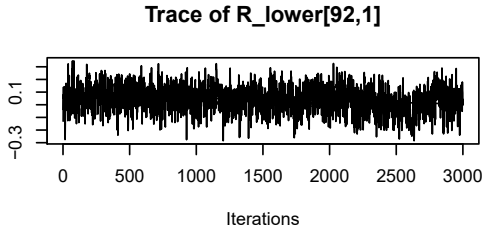

Density of R_lower[89,1]
12
6
0
−0.80
−0.75
−0.70
−0.65
−0.60
N = 3000 Bandwidth = 0.006609

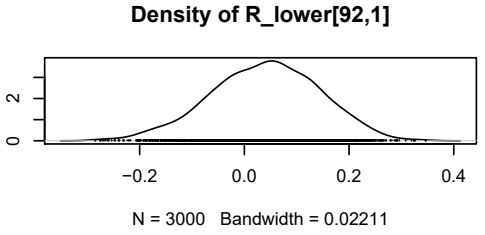

Trace of R_lower[90,1]
−0.30
−0.55
0
500
1000
1500
2000
2500
3000
Iterations

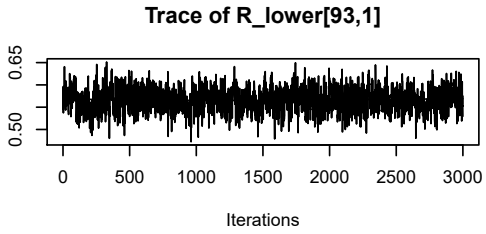

Density of R_lower[90,1]
8
4
0
−0.55
−0.50
−0.45
−0.40
−0.35
−0.30
−0.25
N = 3000 Bandwidth = 0.008198

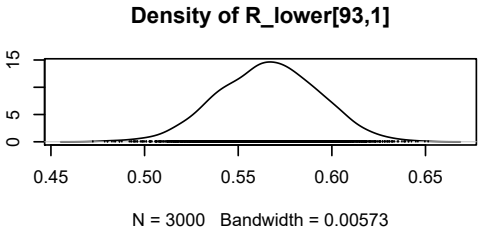

Trace of R_lower[91,1]
0.2
−0.2
0
500
1000
1500
2000
2500
3000
Iterations

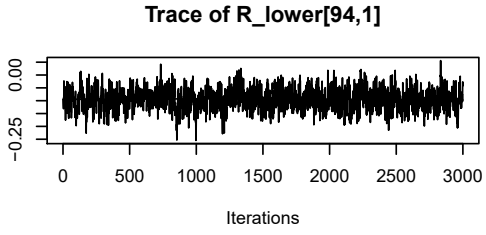

Density of R_lower[91,1]
4
2
0
−0.2
0.0
0.2
0.4
N = 3000 Bandwidth = 0.02144

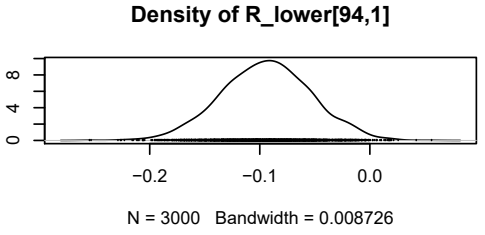

Trace of R_lower[92,1]
0.1
−0.3
0
500
1000
1500
2000
2500
3000
Iterations

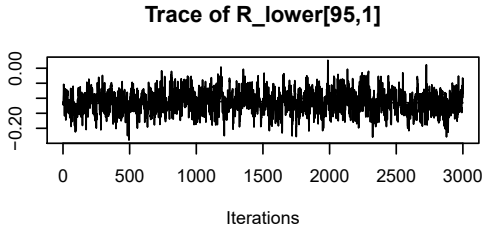

Density of R_lower[92,1]
2
0
−0.2
0.0
0.2
0.4
N = 3000 Bandwidth = 0.02211

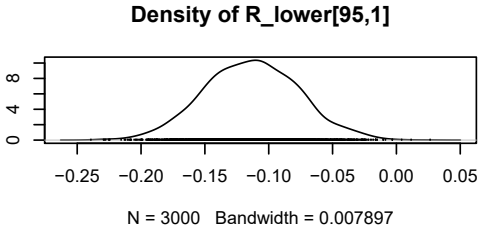

Trace of R_lower[93,1]
0.50
0.65
0
500
1000
1500
2000
2500
3000
Iterations

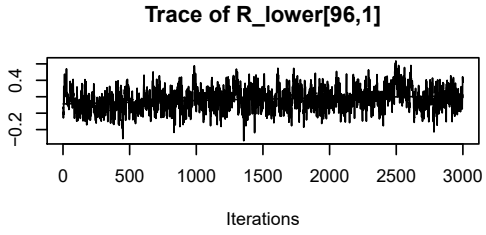

Density of R_lower[93,1]
0
5
15
0.45
0.50
0.55
0.60
0.65
N = 3000 Bandwidth = 0.00573

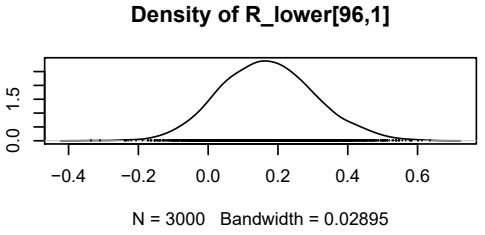

Trace of R_lower[94,1]
−0.25
0.00
0
500
1000
1500
2000
2500
3000
Iterations

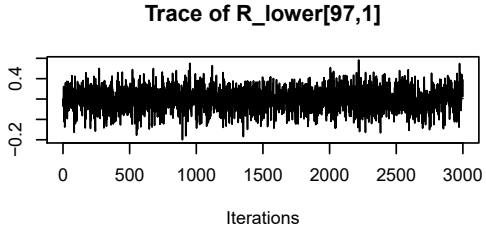

Density of R_lower[94,1]
0
4
8
−0.2
−0.1
0.0
N = 3000 Bandwidth = 0.008726

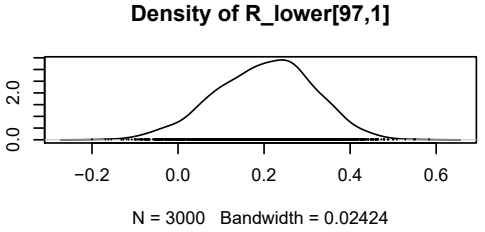

Trace of R_lower[95,1]
−0.20
0.00
0
500
1000
1500
2000
2500
3000
Iterations

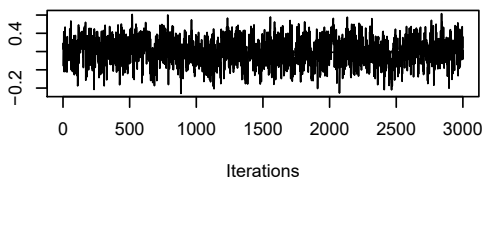

Density of R_lower[95,1]
0
4
8
−0.25
−0.20
−0.15
−0.10
−0.05
0.00
0.05
N = 3000 Bandwidth = 0.007897

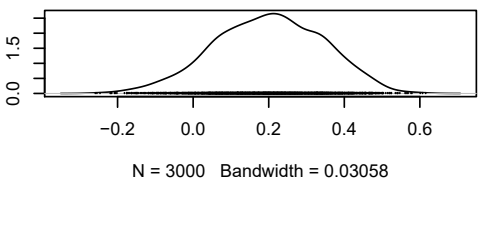

Trace of R_lower[96,1]
−0.2
0.4
0
500
1000
1500
2000
2500
3000
Iterations

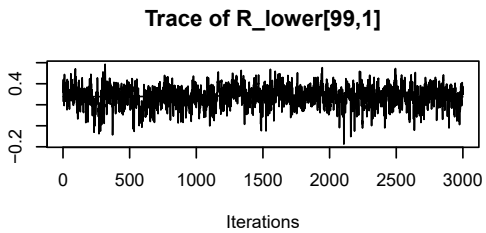

Density of R_lower[96,1]
0.0
1.5
−0.4
−0.2
0.0
0.2
0.4
0.6
N = 3000 Bandwidth = 0.02895

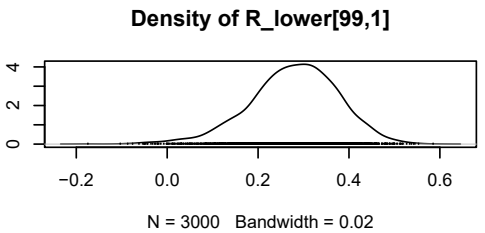

Trace of R_lower[97,1]
−0.2
0.4
0
500
1000
1500
2000
2500
3000
Iterations

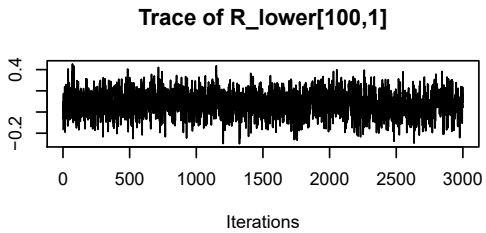

Density of R_lower[97,1]
0.0
2.0
−0.2
0.0
0.2
0.4
0.6
N = 3000 Bandwidth = 0.02424

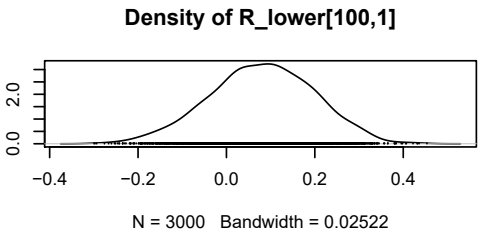

Trace of R_lower[98,1]
−0.2
0.4
0
500
1000
1500
2000
2500
3000
Iterations

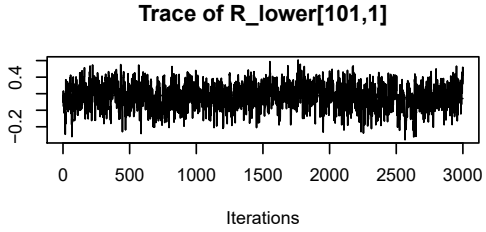

Density of R_lower[98,1]
0.0
1.5
−0.2
0.0
0.2
0.4
0.6
N = 3000 Bandwidth = 0.03058

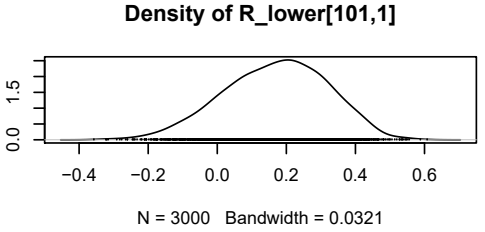

Trace of R_lower[99,1]
−0.2
0.4
0
500
1000
1500
2000
2500
3000
Iterations

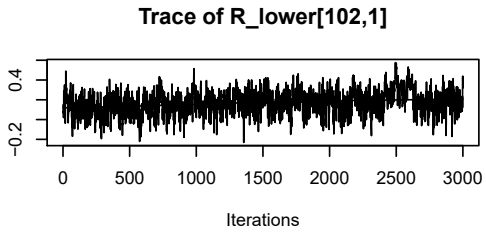

Density of R_lower[99,1]
0
2
4
−0.2
0.0
0.2
0.4
0.6
N = 3000 Bandwidth = 0.02

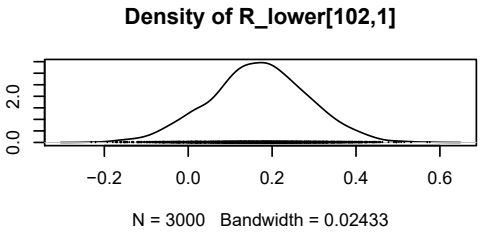

Trace of R_lower[100,1]
−0.2
0.4
0
500
1000
1500
2000
2500
3000
Iterations

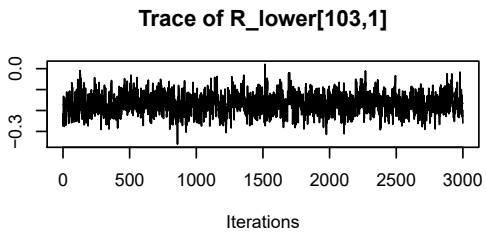

Density of R_lower[100,1]
0.0
2.0
−0.4
−0.2
0.0
0.2
0.4
N = 3000 Bandwidth = 0.02522

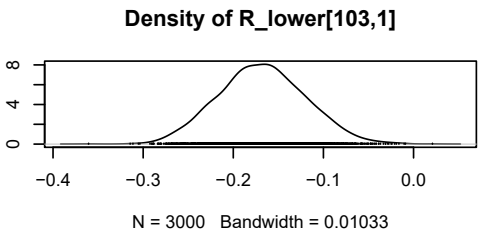

Trace of R_lower[101,1]
−0.2
0.4
0
500
1000
1500
2000
2500
3000
Iterations

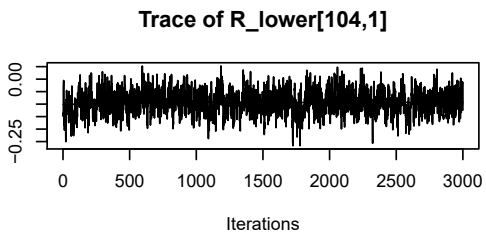

Density of R_lower[101,1]
0.0
1.5
−0.4
−0.2
0.0
0.2
0.4
0.6
N = 3000 Bandwidth = 0.0321

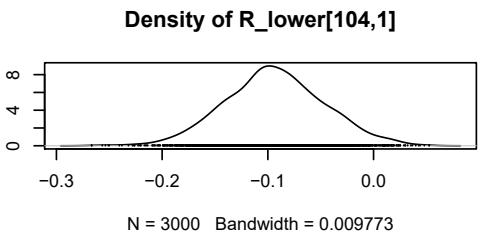

Trace of R_lower[102,1]
−0.2
0.4
0
500
1000
1500
2000
2500
3000
Iterations

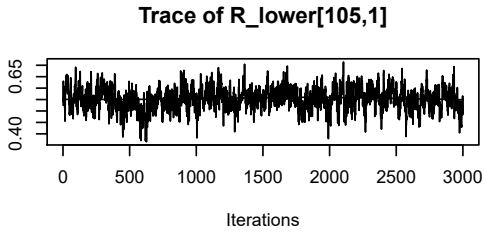

Density of R_lower[102,1]
0.0
2.0
−0.2
0.0
0.2
0.4
0.6
N = 3000 Bandwidth = 0.02433

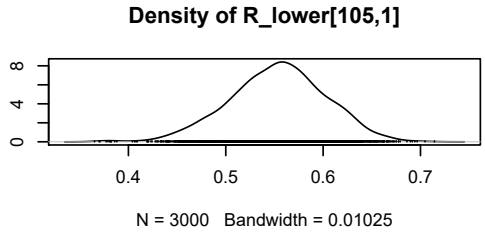

Trace of R_lower[103,1]
−0.3
0.0
0
500
1000
1500
2000
2500
3000
Iterations


Density of R_lower[103,1]
0
4
8
−0.4
−0.3
−0.2
−0.1
0.0
N = 3000 Bandwidth = 0.01033

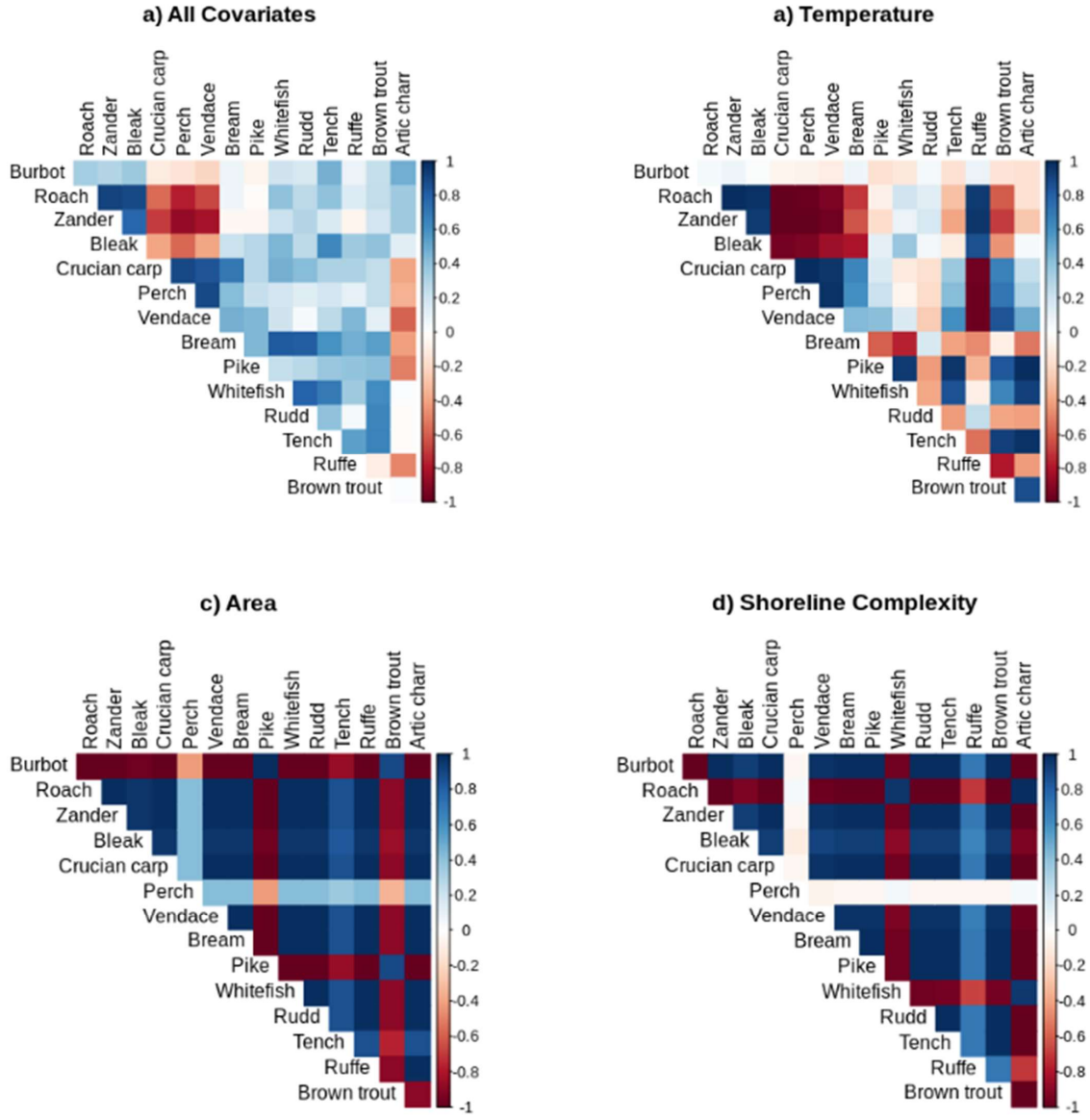

Trace of R_lower[104,1]
−0.25
0.00
0
500
1000
1500
2000
2500
3000
Iterations


Density of R_lower[104,1]
0
4
8
−0.3
−0.2
−0.1
0.0
N = 3000 Bandwidth = 0.009773


Trace of R_lower[105,1]
0.40
0.65
0
500
1000
1500
2000
2500
3000
Iterations


Density of R_lower[105,1]
0
4
8
0.4
0.5
0.6
0.7
N = 3000 Bandwidth = 0.01025

Supplementary Figure S4.1 for *Estimating community-level changes in freshwater species associations over a temperature gradient*



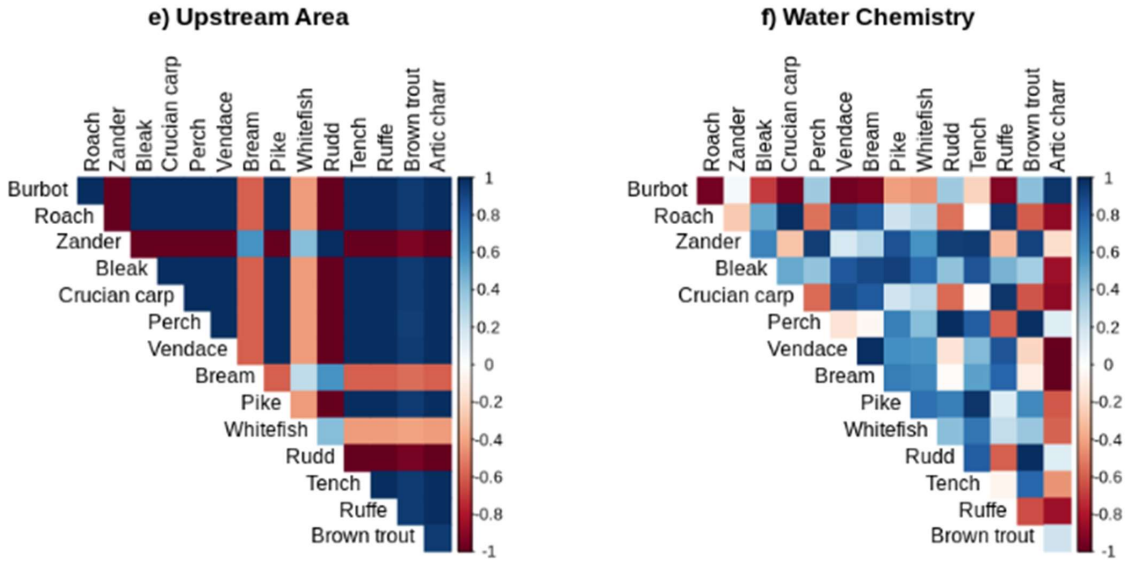


Figure S4.1: Correlation due to shared environmental response of 15 different fish species surveyed across 3308 freshwater lakes in Fennoscandia. Figures show correlation due to response to a) all environmental covariates used in Joint Species Distribution Model, b) average lake surface temperature of warmest quarter of the year, c) lake surface area, d) shoreline complexity, e) upstream area, f) water chemistry (pH and total organic carbon content).

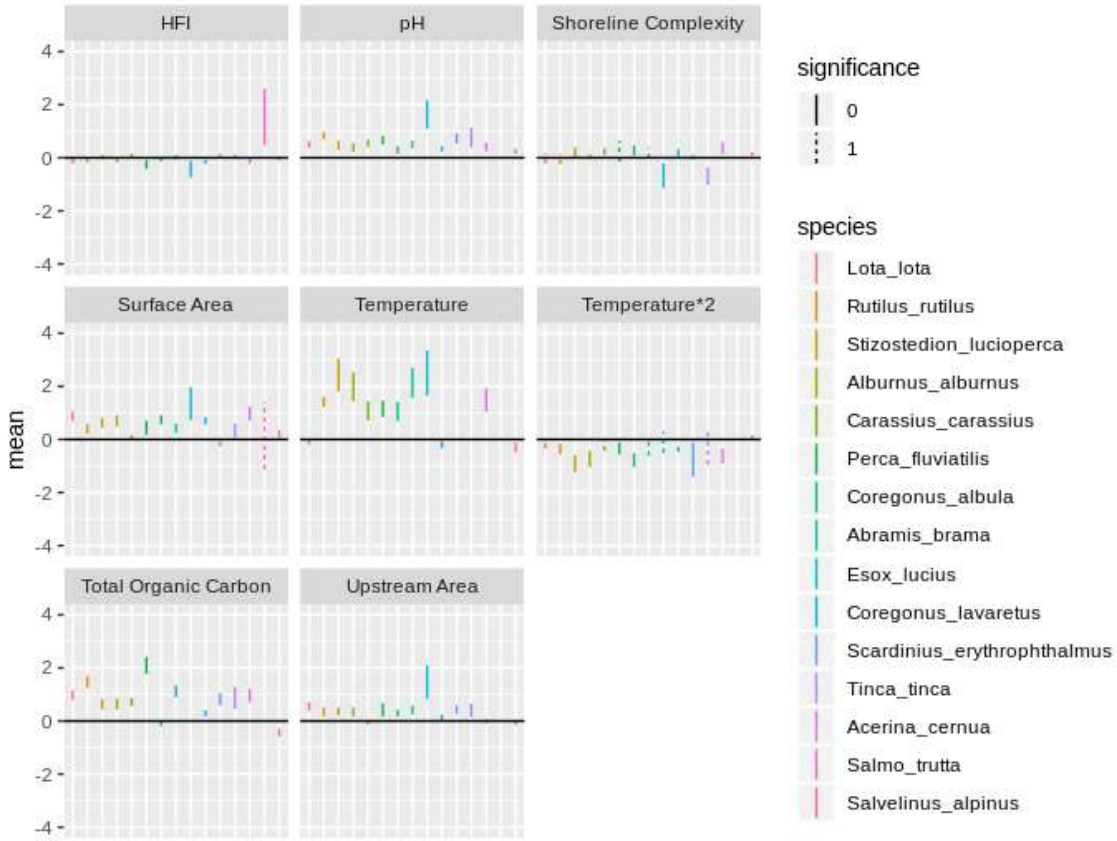
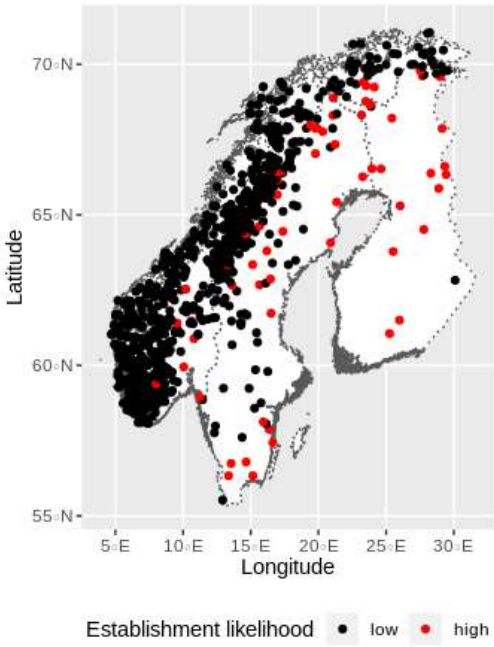
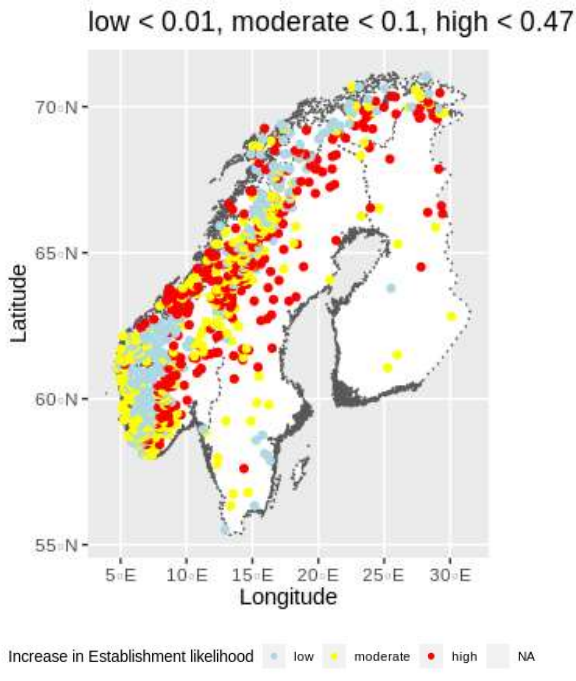


Figure S3.2: 95% credible intervals (CIs) of effect of environmental covariates on presence-absence of 15 different fish species surveyed in 3308 lakes surveyed across Fennoscandia. Dotted lines indicate that CIs intercepted with zero and were therefore not significant.

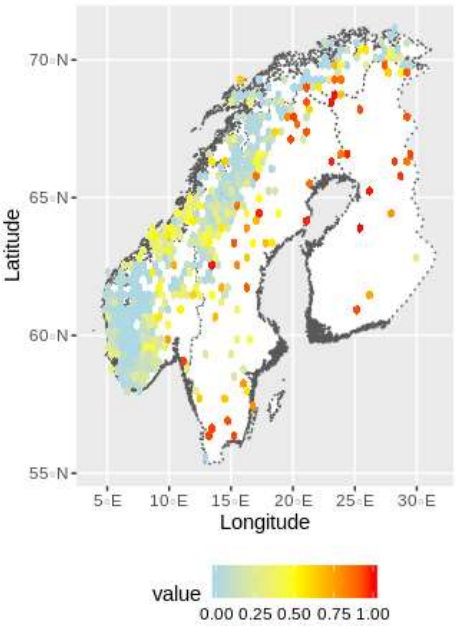
Supplementary Figures S5 for *Estimating community-level changes in freshwater species associations over a temperature gradient*



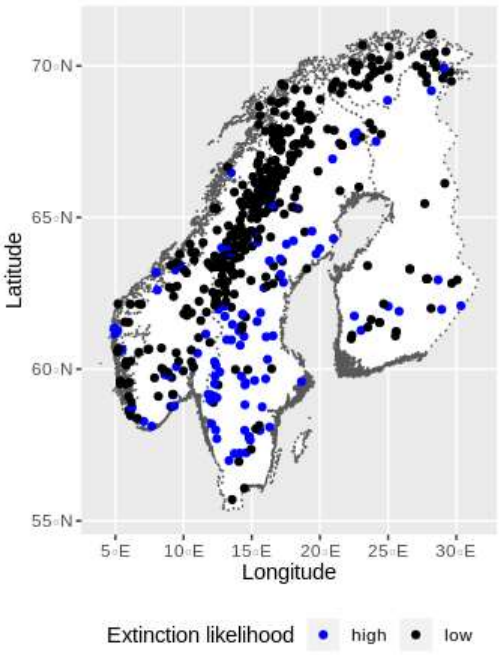
S5.1A: Likelihood of establishment of perch (*Perca fluviatilis*) in lakes surveyed during the 1995 Nordic Freshwater Fish Survey (Tammi et al., 2003) where species was registered as absent. Likelihoods calculated based on a joint species distribution model which accounted for changes in species association over a temperature gradient. Likelihood calculated for a scenario where lake surface temperature has increased by 2 degrees celsius. High likelihood dictated by percentage exceeding the percentage of the 3308 lakes perch was found in during the initial survey. Likelihood of establishment does not account for dispersal barriers.



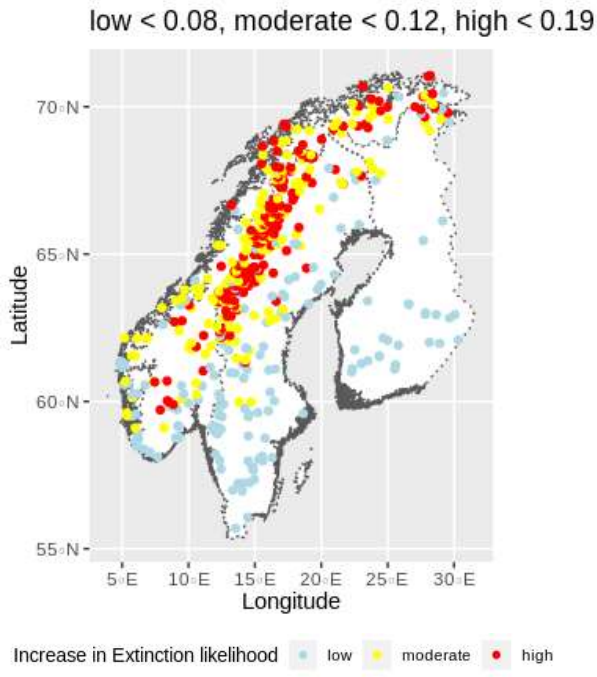
S5.1B: Increase in establishment likelihood of perch (*Perca fluviatilis*) in lakes surveyed during the 1995 Nordic Freshwater Fish Survey (Tammi et al., 2003) given a 2 degree rise in surface temperature. Likelihoods calculated based on a joint species distribution model which accounted for changes in species association over a temperature gradient. Increase in likelihood calculated by subtracting likelihood of establishment calculated for scenario where no warming took place. Likelihood of establishment does not account for dispersal barriers.



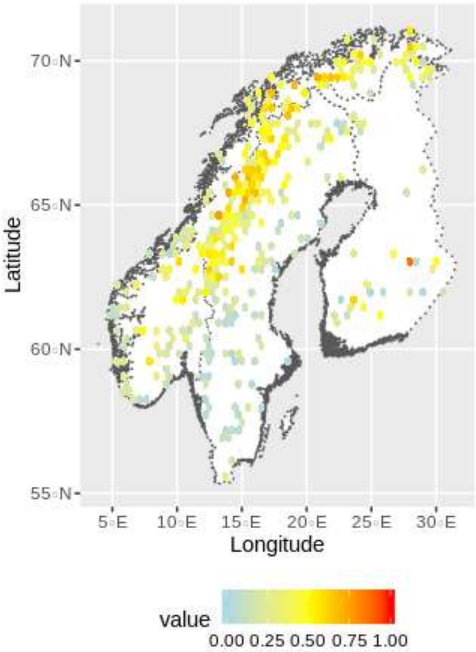
S5.1C: Likelihood of establishment of perch (*Perca fluviatilis*) in lakes surveyed during the 1995 Nordic Freshwater Fish Survey (Tammi et al., 2003) where species was registered as absent. Likelihoods calculated based on a joint species distribution model which accounted for changes in species association over a temperature gradient. Likelihood calculated for a scenario where lake surface temperature has increased by 2 degrees celsius. Likelihood of establishment does not account for dispersal barriers.



S5.2A: Likelihood of local extinction of Arctic charr (*Salvelinus alpinus*) in lakes surveyed during the 1995 Nordic Freshwater Fish Survey (Tammi et al., 2003) where species was registered as present. Likelihoods calculated based on a joint species distribution model which accounted for changes in species association over a temperature gradient. Likelihood calculated for a scenario where lake surface temperature has increased by 2 degrees celsius. Low likelihood dictated by percentage exceeding the percentage of the 3308 lakes Arctic charr was found in during the initial survey. Likelihood of extinction does not account for dispersal barriers preventing other species from establishing in relevant lakes.



S5.2B: Decrease in persistence likelihood of Arctic charr (*Salvelinus alpinus*) in lakes surveyed during the 1995 Nordic Freshwater Fish Survey (Tammi et al., 2003) given a 2 degree rise in surface temperature. Likelihoods calculated based on a joint species distribution model which accounted for changes in species association over a temperature gradient. Decrease in likelihood calculated by subtracting likelihood of persistence calculated for scenario where no warming took place. Likelihood of extinction does not account for dispersal barriers preventing other species from establishing in relevant lakes.



S5.2C: Likelihood of persistence of Arctic charr (*Salvelinus alpinus*) in lakes surveyed during the 1995 Nordic Freshwater Fish Survey (Tammi et al., 2003) where species was registered as absent. Likelihoods calculated based on a joint species distribution model which accounted for changes in species association over a temperature gradient. Likelihood calculated for a scenario where lake surface temperature has increased by 2 degrees celsius. Likelihood of extinction does not account for dispersal barriers preventing other species from establishing in relevant lakes.

Tammi, J., Appelberg, M., Beier, U., Hesthagen, T., Lappalainen, A., & Rask, M. (2003). Fish status survey of Nordic lakes: effects of acidification, eutrophication and stocking activity on present fish species composition. *Ambio*, *32*(2), 98–105.

2.2 Effects of urbanization on pollinator communities and their floral resources

Effects of urbanization on pollinator communities and their floral resources

Elena Albertsen[1] Bert van der Veen[1,2,3] Marie Vestergaard Henriksen[1]

Sølvi Wehn[4] Line Johansen[1]

[1] Norwegian Institute of Bioeconomy Research (NIBIO), NO-1431 Ås, Norway.

[2] Department of Biology, Centre for Biodiversity Dynamics, Norwegian University of Science and Technology, Trondheim, Norway.

[3] Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway.

[4]Multiconsult, 7031 Trondheim, Norway

Abstract

Habitat loss is a major contributor to the decline of pollinators worldwide, with urbanization as an important driver. Our aim in this article was to investigate how plant and pollinator communities are affected by urbanization. We investigated this effect at 28 sites for two high quality habitats (semi-natural grasslands and road verges), and with different levels of urbanization in Trondheim, Norway. To test for the effects of urbanization we modelled the distribution of plant and pollinator communities using Generalized Linear Latent Variable Models. Overall, this analysis revealed that there was an effect of urbanization on both bees and their floral resources, but that effect did not differ between road verge and semi-natural grassland habitats. The distribution of flowering plant species was evenly spread along the two latent variables included in the model, with certain floral resources being more abundant at either ends. Pollinators on the other hand, were more clustered, and certain species were only found to have greater abundance as urbanization increases. This mismatch was somewhat expected given the wider variety of plant groups included in this study. We conclude, that when managed properly, semi-natural grasslands

and road verges have the potential to promote pollinator diversity and abundance both in urban and rural areas.

Key words: Semi-natural grassland, plant-pollinator co-occurrence, urbanization gradient, road verge, Norway, Trondheim, Joint Species Distribution Model

Introduction

A major cause of the worldwide decline in pollinators and their floral resources (Potts et al. 2010) is linked to habitat loss, with urbanization being identified as an important driver (McKinney 2002, Seto et al. 2012). Urbanisation processes include not only the urban sprawl but also agricultural intensification and abandonment of extensive agriculture (Antrop 2004). Urbanisation therefore has an ecological impact on city areas as well as the surrounding agricultural and rural landscapes (Grimm et al. 2008). Urban areas are expected to triple by the year 2030, so it is likely that the process of urbanisation will increase in severity (Seto et al. 2012). It is therefore imperative that we increase our knowledge of how urbanisation affects biodiversity, and ecological communities, in order to develop management and land use planning that can safeguard biodiversity.

In response to agricultural intensification and abandonment there has been a massive loss of semi-natural grasslands, which are among Europe's most species rich habitats (Billeter et al. 2008, Veen et al. 2009). As a result, these semi-natural habitats have become smaller, and have been increasingly fragmented in the landscape (Aune et al. 2018) and are therefore now threatened (Norderhaug and Johansen 2011). Semi-natural grasslands in particular are important habitats for pollinators, due to their high richness of flowering plant species that provide flower resources throughout the summer months, and due to their ability to provide nesting sites (Potts et al. 2003, Kallioniemi et al. 2017, Johansen et al. 2019). Due to the massive loss of semi-natural grasslands in the last century, environmental schemes are now in place that safeguard plant and pollinator communities (Kleijn and Sutherland 2003, Wehn et al. 2018). Few remaining semi-natural grasslands can still be found in both urban and agricultural landscapes, and so a key issue for conservation of biodiversity is to increase the connectivity between these remnant habitats (Öckinger et al.

2009, Krauss et al. 2010, Beninde et al. 2015). New green infrastructure, such as road verges, forest clear cuts and power lines corridors may be high quality habitats for pollinators and plants in both agricultural and urban landscapes (Hovd and Skogen 2005, Cousins 2006, Auestad et al. 2011, Eldegard et al. 2017, Ram et al. 2020, Steinert et al. 2020). These new habitats have the potential to play an important role in the movement of plant and pollinator species between these semi-natural grasslands, thus providing an opportunity to improve connectivity and metapopulation dynamics.

The persistence of high-quality habitats, such as semi-natural grasslands and road verges, is one of the most important factors affecting pollinator and plant species richness (Wenzel et al. 2020). Management that restores, or establishes, high quality habitats within cities and surrounding areas might go a long way in improving the situation for pollinators and their floral resources (Wenzel et al. 2020). However, we do first need to understand how these high-quality habitats are affected by urbanisation, so that suitable conservation schemes for pollinators and their floral resources might be developed, including urban structures.

The evidence for the effects of urbanization on pollinator communities is mixed (Wenzel et al. 2020, Silva et al. 2020). Some studies show that with an increase in agricultural landscapes that are more intensively managed, urban areas have the potential to act as pollinator refugia (Baldock et al. 2015, Hall et al. 2016). Areas with intermediate urbanization often have greater environmental heterogeneity (Winfree et al. 2007, Banaszak-Cibicka et al. 2018), which can facilitate pollinator biodiversity. Within these areas, pollinators can obtain floral resources and nesting sites from semi-natural and natural habitats as well as from gardens and parks (Hinners and Hjelmroos-Koski 2009, Matteson and Langellotto 2011, Garbuzov and Ratnieks 2014, Garbuzov et al. 2015).

The plant-pollinator relationship can be highly asymmetric and nested (Burkle and Alarcón 2011) and to increase our understanding of the effects of urbanisation on biodiversity, it is vital to study communities at several trophic levels including both plants and pollinators simultaneously. Plants and pollinators are dependent on each other for pollination and resources, and so plant species composition is likely to affect the pollinator community (Kearns and Oliveras 2009, Bates et al. 2011, Banaszak-Cibicka and Żmihorski 2012). Many studies have found that species richness between these two groups tends to be highly correlated (Steffan-Dewenter and Tscharntke 2001, Potts et al. 2003, Ebeling et al. 2008, Fründ et al. 2010,

Theodorou et al. 2017), thus, to fully capture the interdependence of plants and their pollinators, effects of urbanisation on both communities needs to be considered.

Factors associated with urbanization, such as turnover in floral resources and increased fragmentation, can favour pollinator species that possess certain traits. Species that are cavity nesting (Cane et al. 2006, Wojcik 2011, Banaszak-Cibicka and Żmihorski 2012, Hinners et al. 2012, Cardoso and Gonçalves 2018), social (Banaszak-Cibicka and Żmihorski 2012, Hinners et al. 2012, Cardoso and Gonçalves 2018), generalist (Bergerot et al. 2010, Banaszak-Cibicka and Żmihorski 2012, Geslin et al. 2013, Wray and Elle 2015), and late emergent (Stelzer et al. 2010, Banaszak-Cibicka and Żmihorski 2012, Wray and Elle 2015) have been argued to benefit from urbanization. Another pollinator trait that is thought to be favoured with increasing urbanization is a larger body size (Banaszak-Cibicka and Żmihorski 2012, Hinners et al. 2012, Martins et al. 2013, Geslin et al. 2013, Merckx et al. 2018). Large-bodied insects have higher mobility, which can be beneficial in fragmented landscapes, however the effect also depends on the sizes of the resource patches (Wenzel et al. 2020).

Our aim with this paper was to investigate how plant and pollinator communities are affected by urbanization within two potentially high-quality habitats: semi-natural grasslands and road verges. To do this, we modelled the distribution of plant and pollinator communities jointly across 28 sites (14 road verges and 14 semi-natural grasslands), which were selected along a gradient of increasing urbanization. The study was conducted in the city of Trondheim, Norway, which has remnant semi-natural grasslands throughout the city, including in densely populated areas. Since road verges, like semi-natural grasslands, are considered to be high-quality habitats for both plants and pollinators (Hovd and Skogen 2005, Cousins 2006, Auestad et al. 2011), we also investigated whether the effect of urbanization on species communities is the same for road verges as for semi-natural grasslands. To help explain the patterns of species distributions we investigate functional plant traits as indicators of the physical growing conditions as well as pollinator dependence and attraction. Functional traits are the key mechanism by which species respond to ecosystem properties, and can therefore provide important insights into community dynamics (de Bello et al. 2010).

Methods

Study sites

We investigated 14 semi-natural grasslands and 14 road verges within the municipality of Trondheim, Norway. This municipality contains approximately 200.000 inhabitants. The semi-natural grasslands were chosen to best represent the range of urbanisation in the landscape and included patches of remnant semi-natural grasslands even in the most populated areas. Only small roads with established vegetation were chosen (single lane tarmac or gravel road wider than 1 meter). The road verge was chosen within 100-500m from the edge of each semi-natural grassland. At this distance, we expect bees to be able to move between the semi-natural and the road verge habitats, to optimally exploit available resources. The 14 semi-natural grasslands all receive extensive management with grazing or/and mowing.

Urbanization gradient

Area resource maps of the study area, municipality of Trondheim (scale 1:5000 (AR5); Ahlstrøm et al. 2014) were provided by the Norwegian Mapping Authority. The area resource classes were categorised in three land cover groups (urban, agriculture, nature; Table 1). A statistical grid for Norway at the resolution of 1 x 1km (Strand and Bloch 2009) was obtained from the Statistics Norway. Using overlay analyses in QGis 2.18.16 we calculated the area of each area resource class in each grid (1x1 km) throughout the municipality. The urbanization gradient was measured as the proportion of settlement and infrastructure versus the proportion of forest, bogs, mires and open land (alpine areas, parks, road verges, lawn, pastures etc.) in each grid cell, i.e. urban versus rural. Rural areas have greater forest land cover and are therefore used to represent the rurality of the site. This measure of urbanizations ranges from -1 to 1 with -1 being purely forested areas and 1 being purely settlements and infrastructure. Each habitat pair (road verge and semi-natural grassland sampled within the same site) will then also have the same measure of urbanization.

Floral resources

All flowering vascular plants except grasses and sedges were registered in 100x1m transects in each semi-natural grassland and road verge. If a transect of 100 meters did not fit within the shape of the semi-natural grassland, the transect was divided in two parallel transects of 50 m with a distance of at least five meters

between them. For each flowering plant species, the number of flowers was counted in five 5x1m plots placed regularly with 20 m distance along the transect. A flower was defined as a floral unit if a medium sized bee can visit without flying. In the study area the main flowering season ranges from early June to early September. Flowers were counted three times during the flowering season in June, July, and August.

Bee community

At each transect, the bee community was sampled in five pan traps three times over the flowering season (in June, July and August) within a week of the floral abundance counts. The traps were placed randomly along the transect but at least 5 meters apart to avoid any interference between them. Each pan trap was made up of three plastic pans (ml) painted in three different UV florescent colors: yellow, blue, and white (Sparvar RAL luminescent spray 1026, 3107, and 3108, respectively) to attract bees with different flower color preference. Pans were placed at the height of the surrounding vegetation to keep them visible to bees. Pans were filled with water and a drop of soap to break the surface tension. Pan traps were left out for approximately 48h and the content of each trap was then filtered from the water and conserved in 70% ethanol. In an attempt to control for environmental effects on sampling, pan traps were only left out when the weather was expected to be warm and sunny, with low windspeeds. This was confirmed when testing for effects of temperature on bee abundance and diversity (see appendix). All bees were identified in the laboratory using relevant identification keys.

Plant traits

The functional traits of the flowering plant species were requested from the databases LEDA (Kleyer et al. 2008), Ecoflora (Fitter and Peat 1994) and Biolflor (Kühn et al. 2004) using the R-package TR8, and were downloaded on October 11th, 2021. The traits selected for this study were used to help explore the distribution of species across the sites. The traits were selected to gain insight into the physical conditions along the urbanization gradient and their dependence on pollination by bees. Plants have commonly been used as environmental indicators and therefore we collected information about the conditions using the Ellenberg indicator values for nitrogen (EIV_N), moisture (EIV_F) and light (EIV_L). Life strategy (C-S-R) reflects these conditions (nitrogen, moisture and light) and was also collected from the database for interpretation.

We investigated the proportion of species with short life spans (annuals) and ruderal life strategies because these traits are commonly associated with urban landscapes and disturbed habitats (Albrecht and Haider 2013, Petersen et al. 2021).

Mating system (selfing, outcrossing and mixed-mating) and type of reproduction (seed or vegetatively) are traits relating to pollinator dependence and were therefore also included to help interpret co-occurrence patterns of plants with pollinators. For flowering plant species that are mainly dependent on pollination by pollinators to reproduce, we used traits relating to traditional pollination syndromes, i.e. floral symmetry and color (Rosas-Guerrero et al. 2014, Dellinger 2020), to assess what pollen vectors it attracts and whether they are specialized.

Statistical analysis

To investigate how plant and pollinator communities are jointly affected by urbanization within semi-natural grasslands and road verges we fitted Generalized Linear Latent Variable Models (GLLVMs, see Niku et al. 2017, 2019). GLLVMs are Joint Species Distribution Models that can alternatively be understood as model-based ordination (similar to e.g (Canonical) Correspondence Analysis), as they are applied here. The dataset used for the model only included species with more than two non-zero observations. Thus, the count of flowers for plants and individuals for bees were used as the (multivariate) response. Jointly analysing flower and pollinator abundance allows us to better predict underlying ecological gradients, and thus species distributions. Specifically, we applied the approach developed by van der Veen et al. (2021) where ecological gradients are modelled by measured and unmeasured components, where the measured components are the predictors. The predictors of this model are the urbanization gradient and a binary variable indicating whether a site was a semi-natural grassland or a road verge. This approach can be understood as a combination of unconstrained and constrained ordination, where the ecological gradient is always optimally represented, but with the benefits of constrained ordination. Alternatively, the model can be understood as a Joint Species Distribution Model with the same rank constraints for the matrix of species predictor slopes as for the residual covariance matrix. The GLLVMs were fitted with a negative-binomial distribution, since after fitting models with a Poisson distribution, residual diagnostics confirmed that distributional assumptions were violated due to overdispersion. We assumed that the negative-binomial

distribution had a different dispersion parameter for plants than for bees, but that it was the same for species within the two groups. We first fitted models without predictors to determine the dimensionality of the data, with 1-5 latent variables. Since the latent variables in these (first) models are unconstrained, this will inform us of the required number of dimensions to optimally represent the co-occurrence patterns in the data. Since the model with two latent variables best fitted the data, we then continued with two latent variables, and fitted six different models, three models with an interaction between the urbanization gradient and the habitat predictor (as to model the slope to the urbanization gradient separately for semi-natural grasslands and roads), and three models with the two predictors as additive effects. We included models with only fixed effects, a single random effect, or random effects for both latent variables. Including the habitat variable equals the hypothesis that the mean abundance of the community differs at roads compared to semi-natural grasslands. From this analysis we can investigate how plant and pollinator communities change with urbanization and whether road verges and semi-natural grasslands are similarly affected.

From the models that all included the additive effects of habitat and urbanization predictors, we considered models within delta two AIC to be equivalent, and then chose the most parsimonious model as the best model (Burnham and Anderson 2002).

Results

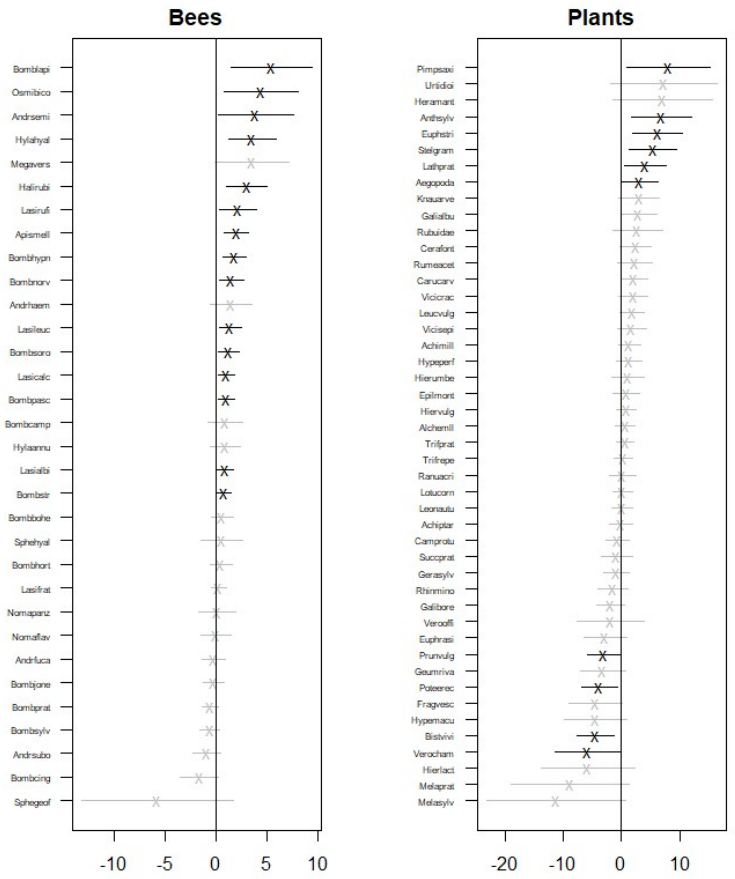
A total of 89 plant species (recorded flowering in the transect) and 42 bee species were registered across all sites. In road verges 67 plant species and 31 bee species were found while there were 69 plant species and 41 bee species in semi-natural grasslands.

Effects of urbanization and habitat type

The best model for flower and bee abundance excluded the interaction between the habitat predictor and the urbanization gradient and included one latent variable with an additional random effect related to it. Only one other model was within delta two AIC, which also included only the additive effect of the predictors, but additionally included a random effect for the second latent variable. Since the standard

deviation for the random effect of this latent variable was close to zero, we determined that it was redundant (van der Veen et al. 2021), so that the second latent variable was sufficiently represented by fixed effects alone.

This joint model on both plants and pollinators showed that urbanization was related to both latent variables, whereas we found no relation with habitat type (LV1: -6.39, CI: -31.45 - 18.67 and LV2: -0.74, CI: -7.26 - 5.79). From the two latent variables, the effect of urbanisation was stronger for LV2 (-1.94, CI: -0.83 - -3.05) than LV1 (0.46, CI: 0.18-0.75), though the confidence interval did not cross zero in either case. For pollinators, the distribution of approximately half of the species along the two latent variables co-occurred in the center of the ordination. The abundances of these species were not affected by urbanization (see Fig. 1). The other half of the bee species were predicted to be most abundant at high degrees of urbanization (Fig. 1). Plants, on the other hand, varied more evenly across the urbanization gradient than the bees (see Fig. 2). This indicates a higher degree of negative co-occurrences and species turnover between floral resources compared to bees along the urbanization gradient. The effects of urbanisation on the abundance of flower resources varied considerably, with the effect being positive for some species and negative for others.



235

236 **Figure 1:** The effects of urbanization in the model on species abundances, including the effects for both
237 latent variables. Effects for bee species are shown on the left and effects for floral resources (flowering plant
238 species) on the right. The solid line indicates zero, the x represents the estimate for each species, and the
239 accompanying solid line is a 95% confidence interval. Species effects of which the confidence interval
240 crossed zero are indicated in grey. See appendix S1: table S1 for the full species names.

241

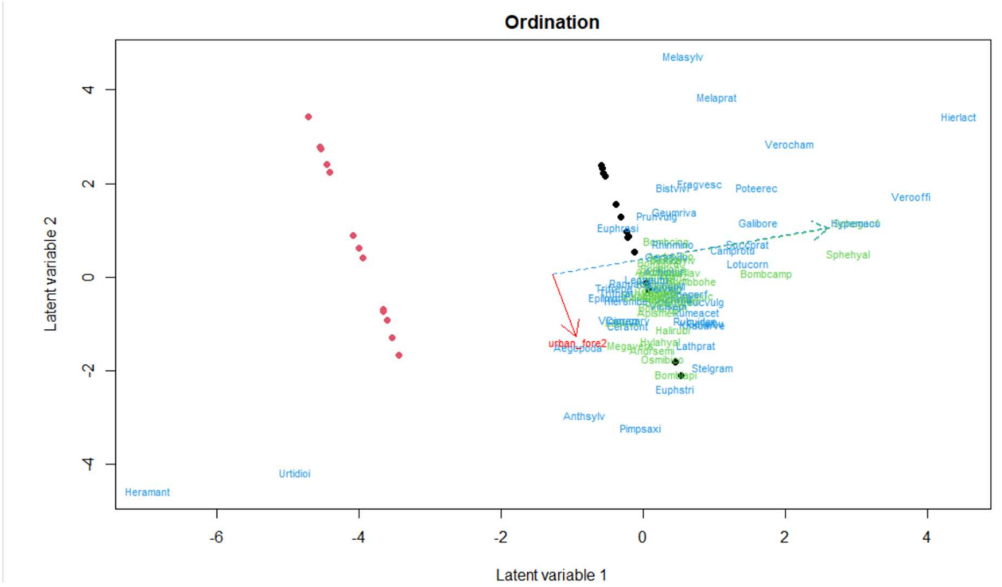


Figure 2: Ordination diagram for the model-based ordination with constrained latent variables. Red dots indicate road verges of sites whereas black dots indicate meadows (though this effect was not significant in the model or ordination). Blue species are plants and green are bees. The red arrow indicates the effect of urbanization, while the effect of the habitat variable in the model was excluded as it was not significant. Higher values of latent variable 2 are indicative of rural areas, and low values of urbanized areas. See appendix S1: table S1 for the full species names.

Plant traits

The broader distribution of flowering plant species away from the bee species was to some degree associated with traits relating to pollinator dependent reproduction (see Fig. 3). Majority of the plants in this study were categorized as insect pollinated but many had alternative strategies for reproduction. Plant species that reproduce solely through outcrossing were equally spread along the latent variables (fig. 3; top right; blue x). However, the outcrossing species that did not overlap with the bee species distributions were also able to reproduce vegetatively (fig. 3; top left; blue plus sign). This was especially true at low urbanization.

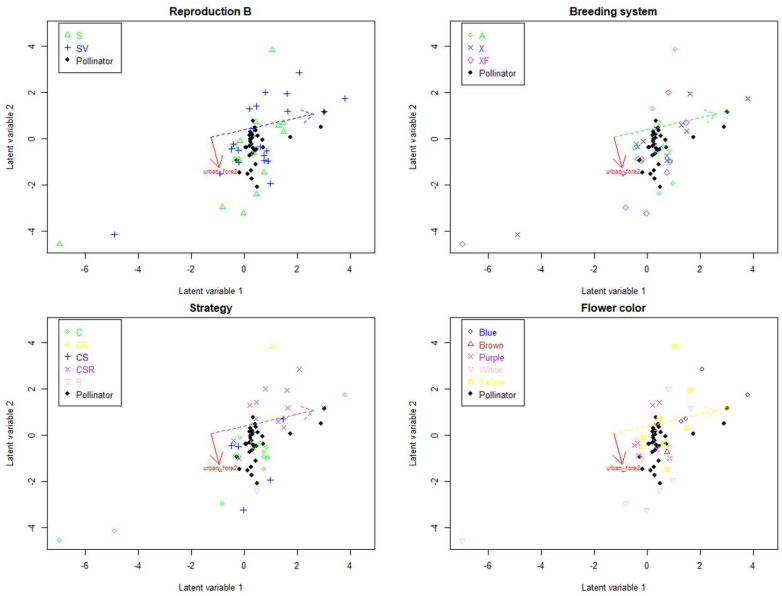


Figure 3: Ordination plot colored by the plant traits; type of reproduction (S = seed, SV = both seed and/or vegetatively), breeding system (A = obligate selfing, X = obligate outcrossing, XF = mixed mating), strategy (C = competitive, CR = competitors/ruderals, CS = competitors/stress-tolerators CSR = competitors/stress-tolerators/ruderals, R = ruderals) and flower color. Bee species are shown as black dots.

Additionally, we found that the pollen vector of these plants may differ as indicated by the flower color. Purple flowers were situated close to the center of the bee distribution, whereas flowering plant species not associated with the bees were either yellow or white. The white flowered species were observed at lower values of both latent variables and yellow flowers more abundant in the higher values.

The greater turnover of flowering plant species was also linked to the physical conditions of the sites (see Fig. 4). Nitrogen affinity seemed to increase with decreasing values of both latent variables and the species with highest affinity (EIV_N values between 7-9) were *Aegopodium podagraria*, *Anthriscus sylvestris*, *Heracleum mantegazzianum*, and *Urtica dioica*. Moisture and light did not show any clear patterns along these latent variables. Overall, we found very few species commonly associated with annual species (6%) and species with a purely ruderal (6%) and ruderal-competitive (5%) life strategies. Species with strategies related to stress-tolerance (s, sr, cs, csr; 54%) and competitive species (c, cr, cs, csr; 89%) make up most of the species recorded. Species associated with stress tolerance, mainly csr, seem to be distributed around

sites of intermediate and lower ranges of urbanization whereas the competitive species were found throughout.

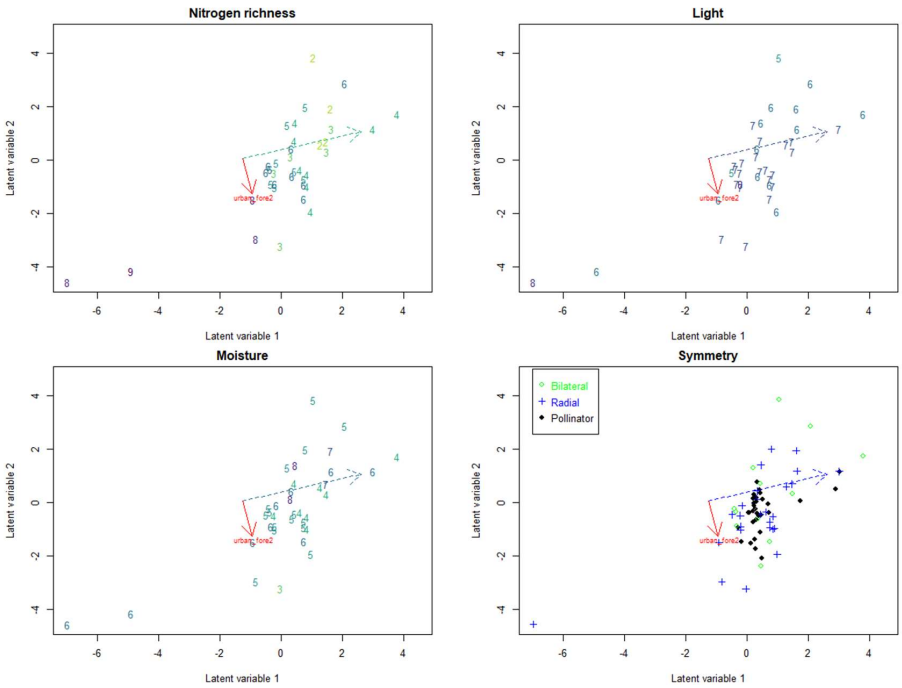


Figure 4: Ordination plot colored by the plant traits related to floral symmetry (Bilateral and radial) and traits indicative of the physical conditions (nitrogen richness = EIV_N, moisture = EIV_F and light = EIV_L) ranging from 1 to 9. Black dots represent the coordinates for bee species.

Discussion

In this article we studied changes in the distribution of bees and floral resources jointly, in order to provide a comprehensive overview of how the pollinator community might change with floral resources in an urban landscape. The inclusion of flower resources is commonly used as predictors in a model for the distribution of bee species (e.g. Potts et al. 2003, Ahrné et al. 2009). The approach used here has the benefit of including both bee and plant species in the same model, as to allow for joint inference based on the relative distances between species coordinates in the ordination. Including abundances of flower resources and bees provides additional information, so that the effect of urbanisation may be determined with a higher degree of

confidence. Overall, the analysis reveals that urbanisation affects the abundance of both bees and floral resources, albeit in different ways (positive for some species and negative for others). We could not conclude that there were any effects of the type of habitat (i.e. semi-natural grassland or road verge) on species abundances. Generally, the distribution of flowering plant species was evenly spread along the two latent variables, whereas pollinators were more clustered. This mismatch was somewhat expected given the wider variety of plant groups included in this study.

The flowering plant species in this study were to some degree insect pollinated, however, many of the recorded flowering plant species had alternative methods for reproduction, i.e., self-pollination, vegetative reproduction or having several pollen vectors. Furthermore, we found that the composition of flowering plant species changed more along the urbanisation gradient than the composition of bee species, with weaker associations between plant and pollinator species at the extremes of the urbanization gradient. Potentially, the distribution of these flowering plant species is not determined by pollinator availability as a result, but rather by other factors. Specifically, at high levels of urbanization, the species in our study that were not strongly associated with bee species were all white and belonged to the family Apiaceae (*Heracleum mantegazzianum, Anthriscus sylvestris and Pimpinella saxifraga*). This family is commonly considered as a generalist and can be pollinated by a wide variety of pollinator groups, such as Coleopterans, Dipterans, Hemipterans, and Hymenopterans (Faegri and van der Pijl 1966, Proctor et al. 1996, Corbet 2006, Zych et al. 2019). Although these species can be pollinated by wild bees (Nielsen et al. 2008, Nichols et al. 2019, Gemeinholzer et al. 2020), our findings suggest that they are unlikely to be their main foraging resource in the landscape. The plant species that were clustered more towards the rural end of the gradient, were less strongly associated with bee species. They were either strong selfers (*Melampyrum sylvaticum, Melampyrum pratense*), fly pollinated (*Potentilla erecta*; Hegland and Totland 2005) or species that commonly reproduce vegetatively (*Veronica chamaedrys,* and *Bistorta vivipara*). And so, the fact that these flowering plant species are more abundant outside the distribution of the bee species, is potentially due to their ability to reproduce, despite low abundance of pollinators in the same places (thus are suspected of having alternative means of reproduction). We could further speculate that this independence from pollinators causes plants to invest less in floral rewards, making them less attractive to pollinators in the community. Given that the deviance of these flowering plant species was mainly explained by lower pollinator dependence, flower color was not as essential in the interpretation of our findings.

Changes in the bee community along the urbanization gradient were driven by solitary bee species that were positively associated with the most urban sites. Various studies, e.g. including Banaszak-Cibicka and Zmihorski (2012), Hinners et al. (2012), Cardoso and Goncalves (2018) have shown that solitary bees are sensitive to urbanization due to their relatively small body size and specialised resource use. However, compared to cities in these studies, Trondheim has a relatively low population density, and an intermediate level of disturbance may explain the positive relationship between bee diversity and urbanisation (Wenzel et al. 2020). Even more likely, the high bee diversity in Trondheim is the positive outcome from a long management history of remnant semi-natural grasslands in the city center, which are mown regularly to maintain biodiversity (Johansen et al. 2019). Higher diversity and abundance of bees suggests better delivery of pollination services to wild plants in road verges and semi-natural grasslands in the city.

In this paper we focused on the effects of urbanisation of plants and pollinators but there are potentially other factors that can be important in explaining this distribution. Other studies investigating the effects of urbanization have considered variables such as connectivity (Fortel et al. 2014), geographical distances between habitats, grazing intensity (Potts et al. 2003), and age and size of the habitat (Ahrné et al. 2009). Patterns of occurrence in plant species were indicative of a nitrogen gradient, where nitrogen seems to decrease with increase in urbanisation, but additionally seems related to unexplained variation in the analysis. The unexplained variation in the analysis could be attributed to the change in nitrogen contents of the soil. Unlike semi-natural grasslands, management at road verges does not include the removal of vegetation after cutting. From this practice, the decomposing vegetation at the road verges could create a more nutrient rich soil depending on the management and explain the pattern of increasing nitrogen affinity across the study sites. Although we could not conclude that the type of habitat was important to explain the distribution of species in this study, we do speculate that nitrogen could be an important driver for the composition of plant species in our study area.

It is imperative to increase our understanding of how the interactions between plants and pollinators are influenced by drivers of global change, such as the urbanisation process. Our study highlights that there is an effect of urbanisation on pollinator communities and their floral resources resulting in a turnover in species present in the city center compared to more rural areas. However, both the city areas and the surrounding cultural landscapes provide high quality habitats of semi-natural grasslands and road verges for pollinators and their floral resources. Therefore, when managed properly, semi-natural grasslands and

road verges have the potential to promote pollinator species diversity and abundance both in urban and rural areas.

Aknowledgements

We thank Arnstein Staverløkk for verifying identified bee species and Julio Morales Can, Per Vesterbukt, Synnøve Nordal Grenne and Annette Bär for their field assistance. BV was supported by a scholarship from the Research Council of Norway (grant number 272408/F40).

Author contributions

MVH, LJ and SW initiated the study. **MVH, BV** and **LJ** collected the data. **BV** performed statistical analysis. **EA** wrote the manuscript with contributions from all authors.

References:

Ahrné, K., J. Bengtsson, and T. Elmqvist. 2009. Bumble Bees (*Bombus spp*) along a Gradient of Increasing Urbanization. PLOS ONE 4:e5574.

Albrecht, H., and S. Haider. 2013. Species diversity and life history traits in calcareous grasslands vary along an urbanization gradient. Biodiversity and Conservation 22:2243–2267.

Antrop, M. 2004. Landscape change and the urbanization process in Europe. Landscape and Urban Planning 67:9–26.

Auestad, I., K. Rydgren, and I. Austad. 2011. Road verges: potential refuges for declining grassland species despite remnant vegetation dynamics. Annales Botanici Fennici 48:289–303.

Aune, S., A. Bryn, and K. A. Hovstad. 2018. Loss of semi-natural grassland in a boreal landscape: impacts of agricultural intensification and abandonment. Journal of Land Use Science 13:375–390.

Baldock, K. C. R., M. A. Goddard, D. M. Hicks, W. E. Kunin, N. Mitschunas, L. M. Osgathorpe, S. G. Potts, K. M. Robertson, A. V. Scott, G. N. Stone, I. P. Vaughan, and J. Memmott. 2015. Where is the UK's pollinator biodiversity? The importance of urban areas for flower-visiting insects. Proceedings of the Royal Society B: Biological Sciences 282.

Banaszak-Cibicka, W., L. Twerd, M. Fliszkiewicz, K. Giejdasz, and A. Langowska. 2018. City parks vs. natural areas - is it possible to preserve a natural level of bee richness and abundance in a city park? Urban Ecosystems 21:599–613.

Banaszak-Cibicka, W., and M. Żmihorski. 2012. Wild bees along an urban gradient: winners and losers. Journal of Insect Conservation 16:331–343.

Bates, A. J., J. P. Sadler, A. J. Fairbrass, S. J. Falk, J. D. Hale, and T. J. Matthews. 2011. Changing bee and hoverfly pollinator assemblages along an urban-rural gradient. PLOS ONE 6:e23459.

de Bello, F., S. Lavorel, S. Díaz, R. Harrington, J. H. C. Cornelissen, R. D. Bardgett, M. P. Berg, P. Cipriotti, C. K. Feld, D. Hering, P. M. da Silva, S. G. Potts, L. Sandin, J. P. Sousa, J. Storkey, D. A. Wardle, and P. A. Harrison. 2010. Towards an assessment of multiple ecosystem processes and services via functional traits. Biodiversity and Conservation 19:2873–2893.

Beninde, J., M. Veith, and A. Hochkirch. 2015. Biodiversity in cities needs space: a meta-analysis of factors determining intra-urban biodiversity variation. Ecology Letters 18:581–592.

Bergerot, B., B. Fontaine, and M. Renard. 2010. Preferences for exotic flowers do not promote urban life in butterflies. Landscape and Urban Planning 96:98–107.

Billeter, R., J. Liira, D. Bailey, R. Bugter, P. Arens, I. Augenstein, S. Aviron, J. Baudry, R. Bukacek, F. Burel, M. Cerny, G. De Blust, R. De Cock, T. Diekötter, H. Dietz, J. Dirksen, C. Dormann, W. Durka, M. Frenzel, R. Hamersky, F. Hendrickx, F. Herzog, S. Klotz, B. Koolstra, A. Lausch, D. Le Coeur, J. P. Maelfait, P. Opdam, M. Roubalova, A. Schermann, N. Schermann, T. Schmidt, O. Schweiger, M. J. M. Smulders, M. Speelmans, P. Simova, J. Verboom, W. K. R. E. Van Wingerden, M. Zobel, and P. J. Edwards. 2008. Indicators for biodiversity in agricultural landscapes: a pan-European study. Journal of Applied Ecology 45:141–150.

Burkle, L. A., and R. Alarcón. 2011. The future of plant–pollinator diversity: Understanding interaction networks across time, space, and global change. American Journal of Botany 98:528–538.

Burnham, K. P., and D. R. Anderson. 2002. Information and likelihood theory: a basis for model selection and inference. Model selection and multimodel inference: a practical information-theoretic approach 2:49–97.

Cane, J. H., R. L. Minckley, L. J. Kervin, A. H. Roulston, and N. M. Williams. 2006. Complex responses within a desert bee guild (Hymenoptera: Apiformes) to urban habitat fragmentation. Wiley Online Library 16:632–644.

Cardoso, M. C., and R. B. Gonçalves. 2018. Reduction by half: the impact on bees of 34 years of urbanization. Urban Ecosystems 21:943–949.

409 Corbet, S. 2006. A typology of pollination systems: implications for crop management and the
410 conservation of wild plants. Pages 315–340 Plant-pollinator interactions. From specialization to
411 generalization. The University of Chicago Press, Chicago.

412 Cousins, S. A. O. 2006. Plant species richness in midfield islets and road verges – The effect of landscape
413 fragmentation. Biological Conservation 127:500–509.

414 Dellinger, A. S. 2020. Pollination syndromes in the 21st century: where do we stand and where may we
415 go? New Phytologist 228:1193–1213.

416 Ebeling, A., A. M. Klein, J. Schumacher, W. W. Weisser, and T. Tscharntke. 2008. How does plant richness
417 affect pollinator richness and temporal stability of flower visits? Oikos 117:1808–1815.

418 Eldegard, K., D. L. Eyitayo, M. H. Lie, and S. R. Moe. 2017. Can powerline clearings be managed to promote
419 insect-pollinated plants and species associated with semi-natural grasslands? Landscape and Urban
420 Planning 167:419–428.

421 Faegri, K., and L. van der Pijl. 1966. The principles of pollination ecology. [1st ed.]. Pergamon Press, New
422 York.

423 Fitter, A. H., and H. J. Peat. 1994. The ecological flora database. The Journal of Ecology 82:415.

424 Fortel, L., M. Henry, L. Guilbaud, A. L. Guirao, M. Kuhlmann, H. Mouret, O. Rollin, and B. E. Vaissière. 2014.
425 Decreasing abundance, increasing diversity and changing structure of the wild bee community
426 (Hymenoptera: Anthophila) along an urbanization gradient. PLOS ONE 9:e104679.

427 Fründ, J., K. E. Linsenmair, and N. Blüthgen. 2010. Pollinator diversity and specialization in relation to
428 flower diversity. Oikos 119:1581–1590.

429 Garbuzov, M., and F. L. W. Ratnieks. 2014. Quantifying variation among garden plants in attractiveness to
430 bees and other flower-visiting insects. Functional Ecology 28:364–374.

431 Garbuzov, M., E. E. W. Samuelson, and F. L. W. Ratnieks. 2015. Survey of insect visitation of ornamental
432 flowers in Southover Grange garden, Lewes, UK. Insect Science 22:700–705.

433 Gemeinholzer, B., J. Reiker, C. M. Müller, and V. Wissemann. 2020. Genotypic and phenotypic distinctness
434 of restored and indigenous populations of *Pimpinella saxifraga* L. 8 or more years after restoration.
435 Plant Biology 22:1092–1101.

436 Geslin, B., B. Gauzens, E. Thébault, and I. Dajoz. 2013. Plant pollinator networks along a gradient of
437 urbanisation. PLoS ONE 8.

438 Grimm, N. B., S. H. Faeth, N. E. Golubiewski, C. L. Redman, J. Wu, X. Bai, and J. M. Briggs. 2008. Global change
439 and the ecology of cities. Science 319:756–760.

440 Hall, D. M., G. R. Camilo, R. K. Tonietto, J. Ollerton, K. Ahrné, M. Arduser, J. S. Ascher, K. C. R. Baldock, R.
441 Fowler, G. Frankie, D. Goulson, B. Gunnarsson, M. E. Hanley, J. I. Jackson, G. Langellotto, D.
442 Lowenstein, E. S. Minor, S. M. Philpott, S. G. Potts, M. H. Sirohi, E. M. Spevak, G. N. Stone, and C. G.
443 Threlfall. 2016. The city as a refuge for insect pollinators. Conservation biology 31:24–29.

444 Hinners, S. J., and M. K. Hjelmroos-Koski. 2009. Receptiveness of foraging wild bees to exotic landscape
445 elements. https://doi.org/10.1674/0003-0031-162.2.253 162:253–265.

446 Hinners, S. J., C. A. Kearns, and C. A. Wessman. 2012. Roles of scale, matrix, and native habitat in supporting
447 a diverse suburban pollinator assemblage. Ecological Applications 22:1923–1935.

448 Hovd, H., and A. Skogen. 2005. Plant species in arable field margins and road verges of central Norway.
449 Agriculture, Ecosystems & Environment 110:257–265.

450 Johansen, L., A. Westin, S. Wehn, A. Iuga, C. M. Ivascu, E. Kallioniemi, and T. Lennartsson. 2019. Traditional
451 semi-natural grassland management with heterogeneous mowing times enhances flower resources
452 for pollinators in agricultural landscapes. Global Ecology and Conservation 18:e00619.

453 Kallioniemi, E., J. Åström, G. M. Rusch, S. Dahle, S. Åström, and J. O. Gjershaug. 2017. Local resources, linear
454 elements and mass-flowering crops determine bumblebee occurrences in moderately intensified

455 farmlands. Agriculture, Ecosystems & Environment 239:90–100.

456 Kearns, C. A., and D. M. Oliveras. 2009. Environmental factors affecting bee diversity in urban and remote
457 grassland plots in Boulder, Colorado. Journal of Insect Conservation 13:655–665.

458 Kleijn, D., and W. J. Sutherland. 2003. How effective are European agri-environment schemes in conserving
459 and promoting biodiversity? Journal of Applied Ecology 40:947–969.

460 Kleyer, M., R. M. Bekker, I. C. Knevel, J. P. Bakker, K. Thompson, M. Sonnenschein, P. Poschlod, J. M. Van
461 Groenendael, L. Klimeš, J. Klimešová, S. Klotz, G. M. Rusch, M. Hermy, D. Adriaens, G. Boedeltje, B.
462 Bossuyt, A. Dannemann, P. Endels, L. Götzenberger, J. G. Hodgson, A. K. Jackel, I. Kühn, D. Kunzmann,
463 W. A. Ozinga, C. Römermann, M. Stadler, J. Schlegelmilch, H. J. Steendam, O. Tackenberg, B. Wilmann,
464 J. H. C. Cornelissen, O. Eriksson, E. Garnier, and B. Peco. 2008. The LEDA Traitbase: a database of life-
465 history traits of the Northwest European flora. Journal of Ecology 96:1266–1274.

466 Krauss, J., R. Bommarco, M. Guardiola, R. K. Heikkinen, A. Helm, M. Kuussaari, R. Lindborg, E. Öckinger, M.
467 Pärtel, J. Pino, J. Pöyry, K. M. Raatikainen, A. Sang, C. Stefanescu, T. Teder, M. Zobel, and I. Steffan-
468 Dewenter. 2010. Habitat fragmentation causes immediate and time-delayed biodiversity loss at
469 different trophic levels. Ecology Letters 13:597–605.

470 Kühn, I., W. Durka, and S. Klotz. 2004. BiolFlor: a new plant-trait database as a tool for plant invasion
471 ecology. Diversity and Distributions 10:363–365.

472 Martins, A., R. Gonçalves, and G. Melo. 2013. Changes in wild bee fauna of a grassland in Brazil reveal
473 negative effects associated with growing urbanization during the last 40 years. Zoologia 30:157–176.

474 Matteson, K. C., and G. A. Langellotto. 2011. Small scale additions of native plants fail to increase beneficial
475 insect richness in urban gardens. Insect Conservation and Diversity 4:89–98.

476 McKinney, M. L. 2002. Urbanization, biodiversity, and conservation. BioScience 52:883–890.

477 Merckx, T., A. Kaiser, and H. Van Dyck. 2018. Increased body size along urbanization gradients at both
478 community and intraspecific level in macro-moths. Global Change Biology 24:3837–3848.

479 Nichols, R. N., D. Goulson, and J. M. Holland. 2019. The best wildflowers for wild bees. Journal of Insect
480 Conservation 23:819–830.

481 Nielsen, C., C. Heimes, and J. Kollmann. 2008. Little evidence for negative effects of an invasive alien plant
482 on pollinator services. Biological Invasions 10:1353–1363.

483 Niku, J., W. Brooks, R. Herliansyah, F. K. C. Hui, S. Taskinen, D. I. Warton, and B. van der Veen. 2017.
484 Package "gllvm." R Project 326.

485 Niku, J., F. K. C. Hui, S. Taskinen, and D. I. Warton. 2019. gllvm: Fast analysis of multivariate abundance
486 data with generalized linear latent variable models in r. Methods in Ecology and Evolution 10:2173–
487 2182.

488 Norderhaug, A., and J. Johansen. 2011. Semi-natural sites and boreal heaths. Page *in* M. Lindgaard, A.,
489 Henriksen, S., Hoem, S., A, & Ødegården, editor. The 2011 Norwegian Red List for Ecosystems and
490 Habitat Types. Norwegian Biodiversity Information Centre.

491 Öckinger, E., M. Franzén, M. Rundlöf, and H. G. Smith. 2009. Mobility-dependent effects on species richness
492 in fragmented landscapes. Basic and Applied Ecology 10:573–578.

493 Petersen, T. K., J. D. M. Speed, V. Grøtan, and G. Austrheim. 2021. Competitors and ruderals go to town:
494 plant community composition and function along an urbanisation gradient. Nordic Journal of Botany
495 39.

496 Potts, S. G., J. C. Biesmeijer, C. Kremen, P. Neumann, O. Schweiger, and W. E. Kunin. 2010. Global pollinator
497 declines: Trends, impacts and drivers. Trends in Ecology and Evolution 25:345–353.

498 Potts, S. G., B. Vulliamy, A. Dafni, G. Ne'eman, and P. Willmer. 2003. Linking bees and flowers: How do floral
499 communities structure pollinator communities? Ecology 84:2628–2642.

500 Proctor, M., P. Yeo, and A. Lack. 1996. The natural history of pollination. HarperCollins Publishers, London.

Ram, D., Å. Lindström, L. B. Pettersson, and P. Caplat. 2020. Forest clear-cuts as habitat for farmland birds and butterflies. Forest Ecology and Management 473:118239.

Rosas-Guerrero, V., R. Aguilar, S. Martén-Rodríguez, L. Ashworth, M. Lopezaraiza-Mikel, J. M. Bastida, and M. Quesada. 2014. A quantitative review of pollination syndromes: do floral traits predict effective pollinators? Ecology Letters 17:388–400.

Seto, K. C., B. Güneralp, and L. R. Hutyra. 2012. Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. Proceedings of the National Academy of Sciences of the United States of America 109:16083–16088.

Silva, J. L. S., M. T. P. de Oliveira, O. Cruz-Neto, M. Tabarelli, and A. V. Lopes. 2020. Plant–pollinator interactions in urban ecosystems worldwide: A comprehensive review including research funding and policy actions. Ambio 2020 50:4 50:884–900.

Steffan-Dewenter, I., and T. Tscharntke. 2001. Succession of bee communities on fallows. Ecography 24:83–93.

Steinert, M., K. Eldegard, M. A. K. Sydenham, and S. R. Moe. 2020. Bumble bee communities in power-line clearings: Effects of experimental management practices. Insect Conservation and Diversity:icad.12463.

Stelzer, R. J., L. Chittka, M. Carlton, and T. C. Ings. 2010. Winter active bumblebees (*Bombus terrestris*) achieve high foraging rates in urban Britain. PLoS ONE 5.

Theodorou, P., K. Albig, R. Radzevičiūtė, J. Settele, O. Schweiger, T. E. Murray, and R. J. Paxton. 2017. The structure of flower visitor networks in relation to pollination across an agricultural to urban gradient. Functional Ecology 31:838–847.

van der Veen, B., F. Hui, K. Hovstad, and Rb. O'Hara. 2021. Model-based ordination with constrained latent variables. bioRxiv.

Veen, P., R. Jefferson, J. De Smidt, and J. Van Der Straaten. 2009. Grasslands in Europe: of high nature value. KNNV publishing.

Wehn, S., R. Burton, M. Riley, L. Johansen, K. A. Hovstad, and K. Rønningen. 2018. Adaptive biodiversity management of semi-natural hay meadows: The case of West-Norway. Land Use Policy 72:259–269.

Wenzel, A., I. Grass, V. V. Belavadi, and T. Tscharntke. 2020. How urbanization is driving pollinator diversity and pollination – A systematic review. Biological Conservation 241:108321.

Winfree, R., T. Griswold, and C. Kremen. 2007. Effect of human disturbance on bee communities in a forested ecosystem. Conservation Biology 21:213–223.

Wojcik, V. A. 2011. Bees (Hymenoptera: Apoidea) utilizing Tecoma stans (L.) Juss. ex Kunth (Bignoniaceae) in urban landscapes: A comparison of occurrence patterns and community composition in three cities in northwestern Costa Rica. Journal of the Kansas Entomological Society 84:197–208.

Wray, J. C., and E. Elle. 2015. Flowering phenology and nesting resources influence pollinator community composition in a fragmented ecosystem. Landscape Ecology 30:261–272.

Zych, M., R. R. Junker, M. Nepi, M. Stpiczyńska, B. Stolarska, and K. Roguz. 2019. Spatiotemporal variation in the pollination systems of a supergeneralist plant: is *Angelica sylvestris* (Apiaceae) locally adapted to its most effective pollinators? Annals of Botany 123:415–428.

Appendix S1

Temperature can affect bee activity and so the number of bees caught within pan trap may vary as a result of this. To account for these effects, we placed temperature loggers at each transect for the same duration as the pan traps. To test for any effects, we analysed the data using a generalized mixed effect model with either bee abundance or bee species richness as the response and temperature, flower abundance and number of flowering species as the explanatory variables. This model also accounted for the time that the survey was completed and the nested sampling design (two habitats within each site). Since flower abundance and richness can also affect bee activity, we included flowering data to control for any confounding effects. When accounting for flowering, temperature did not have a significant effect on bee abundance and species richness in the pan traps and was therefore not used as an explanatory variable in the main analysis.

Table S1: Abbreviated and scientific species names.

Abbreviated	Full name
Apismell	Apis mellifera
Bombcamp	Bombus campestris
Bombhort	Bombus hortorum
Bombpasc	Bombus pascuorum
Bombprat	Bombus pratorum
Bombstr	Bombus s. str.
Bombsoro	Bombus soroeensis
Bombsylv	Bombus sylvestris
Hylaannu	Hylaeus annulatus
Lasialbi	Lasioglossum albipes
Lasicalc	Lasioglossum calceatum
Lasifrat	Lasioglossum fratellum
Lasileuc	Lasioglossum leucopus
Lasirufi	Lasioglossum rufitarse
Nomaflav	Nomada flavoguttata
Andrhaem	Andrena haemorrhoa
Bombhypn	Bombus hypnorum
Andrfuca	Andrena fucata
Bombbohe	Bombus bohemicus
Bombcing	Bombus cingulatus
Bombjone	Bombus jonellus
Bombnorv	Bombus norvegicus

Sphehyal	Sphecodes hyalinatus
Hylahyal	Hylaeus hyalinatus
Sphegeof	Sphecodes geoffrellus
Andrsemi	Andrena semilaevis
Bomblapi	Bombus lapidarius
Halirubi	Halictus rubicundus
Megavers	Megachile versicolor
Osmibico	Osmia bicornis
Andrsubo	Andrena subopaca
Nomapanz	Nomada panzeri
Achimill	Achillea millefolium
Achiptar	Achillea ptarmica
Alchemll	Alchemilla
Anthsylv	Anthriscus sylvestris
Carucarv	Carum carvi
Epilmont	Epilobium montanum
Galialbu	Galium album
Gerasylv	Geranium sylvaticum
Geumriva	Geum rivale
Heramant	Heracleum mantegazzianum
Hypemacu	Hypericum maculatum
Hypeperf	Hypericum perforatum
Leonautu	Leontodon autumnalis
Poteerec	Potentilla erecta
Ranuacri	Ranunculus acris
Rhinmino	Rhinanthus minor
Rubuidae	Rubus idaeus
Rumeacet	Rumex acetosa
Stelgram	Stellaria graminea
Trifprat	Trifolium pratense
Trifrepe	Trifolium repens
Verooffi	Veronica officinalis
Vicisepi	Vicia sepium
Aegopoda	Aegopodium podagraria
Camprotu	Campanula rotundifolia
Fragvesc	Fragaria vesca
Galibore	Galium boreale
Hiervulg	Hieracium vulgata
Knauarve	Knautia arvensis
Lathprat	Lathyrus pratensis
Leucvulg	Leucanthemum vulgare
Prunvulg	Prunella vulgaris
Urtidioi	Urtica dioica
Vicicrac	Vicia cracca
Hierlact	Hieracium lactucella
Lotucorn	Lotus corniculatus

Melaprat	Melampyrum pratense
Succprat	Succisa pratensis
Verocham	Veronica chamaedrys
Cerafont	Cerastium fontanum
Hierumbe	Hieracium umbellatum
Bistvivi	Bistorta vivipara
Euphrasi	Euphrasia
Pimpsaxi	Pimpinella saxifraga
Melasylv	Melampyrum sylvaticum
Euphstri	Euphrasia stricta

557

3

Conquering the universe one GLLVM at a time

3.1 Next generation ordination with Generalized Linear Latent Variable Models

Next generation ordination with Generalized Linear Latent Variable Models

Bert van der Veen[123] Knut A. Hovstad[34] Francis K.C. Hui[5]
Robert B. O'Hara[23]

[1]Department of Landscape and Biodiversity, Norwegian Institute of Bioeconomy research, Trondheim, Norway
[2]Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway
[3]Centre of Biodiversity Dynamics, Norwegian University of Science and Technology, Trondheim, Norway
[4]The Norwegian Biodiversity Information Centre, Trondheim, Norway
[5]Research School of Finance, Actuarial Studies and Statistics, The Australian National University, Canberra, Australia

Abstract

Techniques for the dimension reduction of multivariate datasets are a commonly used tool in ecology. By reducing dimensions, it becomes more straightforward to explore the latent processes that underlay ecological data. Ordination methods such as Correspondence Analysis and non-metric Multidimensional Scaling are well studied, and frequently applied in ecology. However, in recent years newly developed ordination methods using Generalized Linear Latent Variable Models have increased in popularity. Model-based ordination is, in contrast to classical ordination, a more flexible and transparent framework for dimension reduction. The model-based framework allows users to include random effects in the application of ordination and to validate any assumptions that are made during the analysis. This synthesis provides an overview of the similarities between popular classical ordination methods in ecology and model-based ordination, while arguing that Generalized Linear Latent Variable Models are the next generation of ordination methods.

Introduction

Many multivariate datasets in ecology can be considered as generated by some latent process (Austin 1985; ter Braak & Prentice 1988). Techniques that reduce dimensions of multivariate datasets attempt to capture that process as best as possible in few dimensions. Each dimension then represents an aspect of the process, in essence a latent variable (ter Braak & Prentice 1988), to which meaning is attributed using the relative distance of coordinates for rows and columns to each other in the multidimensional space (Gower 1966; Gabriel 1971). Reducing dimensions is useful for high dimensional data, as it allows researchers to better capture and visualize patterns in their data, or to perform inference more straightforwardly.

In ecology, dimension reduction techniques for multivariate datasets are termed "ordination" (Goodall 1954), and are used to analyse abundances or binary data that are generated by the environment at a collection of sites. Each aspect of the environment is referred to as an environmental gradient, and the combination of multiple environmental gradients is referred to as a "complex ecological gradient" (Whittaker 1967; Halvorsen 2012). It is this interpretation to which ordination methods owe their popularity, since unmeasured ecological gradients can be interpreted as a latent variable, so that ordination is a type of latent variable modelling, where the species responses are then modelled as a linear combination of latent variables.

A geometric interpretation of ordination methods follows from the view that they can be understood as methods to summarize a matrix (Jongman *et al.* 1995), e.g. as in Kidziński *et al.* (2021). Unsurprisingly, techniques for matrix decomposition such as an eigendecomposition or singular value decomposition underlie many ordination techniques such as Principal Component Analysis (Pearson 1901), but also Correspondence Analysis (CA, Hirschfeld 1935) and Principal Coordinate Analysis (PCoA, Gower 1966; Anderson & Willis 2003).

Species and sites, corresponding to the columns and rows of a multivariate dataset, are positioned by ordination methods in a lower dimensional space by their dissimilarity. The ordination space can then be visually inspected using a biplot. Gabriel (1971) developed the biplot as a visual display for application to high dimensional data, specifically in the context of PCA. Since its development, the biplot has been applied to many ordination methods (ter Braak & Looman 1994; Niku *et al.* 2019; Hawinkel *et al.* 2019).

Two schools of ordination methods each attempt to describe patterns in the data based on conceptually different frameworks (Roberts 2020): the distance-based school which includes methods such as Principal Coordinate Analysis (PCoA, Gower 1966) and Non-metric Multidimensional Scaling (NMDS, Kruskal 1964a,b), and the model-based school (ter Braak 1985; Walker & Jackson 2011; Hui *et al.* 2015). Recent years have seen a push in the development of model-based ordination methods, in the form of row-column interaction models (Yee & Hadi 2014; Hawinkel *et al.* 2019) and with Generalized Linear Latent Variable

Models (GLLVMs, Warton *et al.* 2015a). PCA and CA can be considered to have a foot in both worlds, since those methods can be connected to a distance measure (Mardia *et al.* 1980 p. 405; Legendre & Legendre 2012 pp. 466–467), but at the same time have an (implicit) statistical model that they relate to (Jongman *et al.* 1995).

In this article, we describe some popular ordination methods, and summarize some of their properties, in order to demonstrate that the GLLVM framework can be considered as a new, more advanced and flexible framework for next generation ordination methods. GLLVMs are a modelling framework that builds on the ideas of Generalized Linear Models (Nelder & Wedderburn 1972) and Generalized Linear Mixed-effects Models (Bolker *et al.* 2009), but facilitate a latent variable interpretation, so that researchers are able to reduce dimensions of ecological data in a more model-oriented, and more statistically rigorous, manner. In various articles, GLLVMs have been shown to retrieve similar ordinations, or even outperform, classical ordination methods (Hui *et al.* 2015; Popovic *et al.* 2019; Jupke & Schäfer 2020; van der Veen *et al.* 2021a; but also see Roberts 2017; and Hawinkel *et al.* 2019). All methods described in this article are a type of (model-based) (residual) ordination with (un)constrained latent variables, but for different data types. In the first section, we start by describing PCA in order to explain the purpose of ordination in more detail, since it is the oldest, most well known, and most analytically accessible ordination method, at least when data are well behaved. In the second section we will discuss popular ordination methods in ecology, most notably NMDS and (D)CA. Thirdly, we explain the model-based ordination framework for unconstrained ordination. Thereafter we discuss ordination methods that constrain latent variables using an additional matrix of predictors, using explicit statistical models. Finally, we conclude this article with a discussion of the potential and flexibility of the GLLVM framework for the ordination of ecological communities.

Introducing ordination

PCA was first applied in ecology by Goodall (1954) in order to study species co-occurrence patterns in a plant community. Nowadays PCA is infrequently used for the analysis of ecological communities due to its "flaws" (Swan 1970; Legendre & Legendre 2012).

For a matrix $\boldsymbol{Y}$ of observations in $j = 1 \dots p$ columns with $i = 1 \dots n$ rows, with column means $\bar{\boldsymbol{y}}_j$ PCA rotates the sample covariance matrix of the data to a coordinate system with orthogonal axes using an eigendecomposition:

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{V}\text{diag}(\boldsymbol{\lambda})\boldsymbol{V}^\top \tag{1}$$

where $\hat{\boldsymbol{\Sigma}} = (\boldsymbol{Y} - \bar{\boldsymbol{y}}_j)^\top(\boldsymbol{Y} - \bar{\boldsymbol{y}}_j)(n-1)^{-1}$, where $\boldsymbol{V}$ is a $p \times p$ unitary matrix whose columns are referred to as loadings or eigenvectors (i.e., $\boldsymbol{v}_j$ the coordinates for each column in the data as part of the multidimensional

space) which are orthonormal (i.e. orthogonal and have unit length, so that we have the constraint $\boldsymbol{v}_j^\top \boldsymbol{v}_j = 1$), and where $\boldsymbol{\lambda} = \{1 \geq \lambda_1 > \dots \lambda_d\}$ are eigenvalues which are ordered in a decreasing manner (Mardia *et al.* 1980). The order constraint on the eigenvalues emphasises that the first axis explains most variance, the second axis thereafter and so forth, so that PCA can be thought of as the method that finds a p dimensional hyperplane with maximum variance order constraint, and with additional constraints on the loadings. Mardia *et al.* (1980) (pp. 229) summarizes some properties of PCA for maximum likelihood estimation of normal responses but also see Lynn *et al.* (1995) and Lynn & McCulloch (2000).

Inspecting the ordination

When inspecting a PCA through a biplot, the eigenvectors for d specific (by the user chosen, but generally the first two) dimensions are plotted as arrows that start at the origin of Principal Component (PC) axes and extend to the column coordinate calculated by PCA. The angle of an arrow to the PC axes represent the relation of that column to the PC axes, and the length of the arrow reflects the importance of a PC axis in explaining that column.

For a complete overview of patterns in the data, coordinates of rows (also known as row scores) for a PCA can be retrieved by using a singular value decomposition:

$$\bar{\boldsymbol{Y}} = \boldsymbol{U}\text{diag}(\boldsymbol{\lambda}^{0.5})\boldsymbol{V}^\top, \tag{2}$$

for the orthonormal matrix $\boldsymbol{U}$ that holds all row scores, also known as the left singular vectors (i.e. the eigenvectors of the eigendecomposition for $(\boldsymbol{Y} - \bar{\boldsymbol{y}}_j)(\boldsymbol{Y} - \bar{\boldsymbol{y}}_j)^\top (n-1)^{-1}$), $\boldsymbol{\lambda}^{0.5}$ are singular values (i.e. the square root of the eigenvalues, representing the scale of the eigenvectors), and where the right singular vectors are the same as the eigenvectors above (Legendre & Legendre 2012 pp. 461–462). Note that the right singular vectors can be calculated using the left singular vectors, since $\boldsymbol{U} = \bar{\boldsymbol{Y}}\boldsymbol{V}^\top\text{diag}(\boldsymbol{\lambda}^{-0.5})$, so that both sets of scores can be retrieved with a single eigendecomposition.

Naturally, in the context of a latent variable model, the coordinate system can be rotated or scaled in any other way while retaining a similar interpretation, but the maximum variance rotation of PCA has an appealing property for applied research. With any other rotation it might be difficult to determine which of the dimensions should be chosen to draw a low-dimensional plot (Gabriel 1971) in order to explore patterns in the subjected matrix. Here, since the eigenvectors explain the maximum possible variation each after each other, it is now possible to simply take the first few dimensions that account for most of the variation in the matrix, for example using a screeplot (Cattell 1966).

Distance-based view

An alternative way of considering PCA, instead of based on the sample covariance matrix, is by the distances between the coordinates on a hyperplane. The distance between the coordinates is Euclidean in nature (Legendre & Legendre 2012 p. 433; Greenacre 2017), which can be seen by noting that the PCs are a linear transformation of the original columns in the data, and by noting that the sample covariance matrix is proportional to the (element-wise squared) Euclidean distance matrix (Mardia *et al.* 1980 pp. 404–405; Hastie *et al.* 2016 p. 671). For a column-centred matrix of observations $\bar{\boldsymbol{Y}}$ with entries $\bar{y}_{ij}$ for row $i = 1 \dots n$ and column $j, m = 1 \dots p$ we can write $\hat{\Sigma}_{jm} = \sigma_j^2 + \sigma_m^2 - 2\sigma_{jm}$ and set $\sigma_j^2 = \sum_{i=1}^{n} \bar{y}_{ij}^2$, $\sigma_m^2 = \sum_{i=1}^{n} \bar{y}_{im}^2$, and $\sigma_{jm} = \sum_{i=1}^{n} \bar{y}_{ij}\bar{y}_{im}$ as in the squared Euclidean distance. Then, for $j = m$ i.e. the diagonal entries, we have $\hat{\Sigma}_{jj} = 0$ as expected. As such, the eigendecomposition of the two matrices is identical. It is this view that connects PCA to the other distance-based methods discussed in this article.

Model-based view

PCA can alternatively be re-formulated as a method for multivariate linear regression (Jong & Kotz 1999), and in general can be considered a multivariate equivalent of linear regression (ter Braak & Prentice 1988; Lynn *et al.* 1995; Jongman *et al.* 1995; Lynn & McCulloch 2000), so that PCA performs best under the assumption of multivariate normality. If we remove the $(p - d)$ eigenvectors with the smallest eigenvalues, PCA approximates the data with the following linear latent variable model:

$$y_{ij} = \boldsymbol{z}_i^\top \boldsymbol{\gamma}_j, \tag{3}$$

where $\boldsymbol{z}_i$ is a vector of d scores for row i treated as fixed effect, and $\boldsymbol{\gamma}_j$ is a vector of d column loadings for column j. Omitting the first d eigenvectors that explain comparably little variation in the response implies a residual error ϵ_{ij}. Since the row scores, column loadings, and residual are linear combinations of the data, they too follow a normal distribution if the responses are normally distributed.

Factor Analysis (FA, Spearman 1904) is a method akin to PCA, that was developed for application in psychology. FA also fits a linear latent variable model, but with an explicit error term:

$$y_{ij} = \beta_{0j} + \boldsymbol{z}_i^\top \boldsymbol{\gamma}_j + \epsilon_{ij}, \tag{4}$$

where $\boldsymbol{z}_i \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I})$ is a vector of d row scores, and $\boldsymbol{\gamma}_j$ is a vector of d column loadings which are similar to the eigenvectors for PCA (Bartholomew 2011 p. 61), and where $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\boldsymbol{0}, \mathrm{diag}(\boldsymbol{\sigma}_j^2))$. In PCA the data are column-centred by the sample mean, corresponding to the maximum likelihood estimator for the intercept in

the factor analysis model, if the data is normally distributed. From the same perspective, PCA finds latent variables under the assumption that $\boldsymbol{\sigma}_j = 0$, i.e. PCA assumes the data can be perfectly represented by the latent variables alone, so that the solutions can be expected to be similar when the error variances are small (Bartholomew 2011 p. 61), so that PC axes are contaminated with residual error otherwise.

FA finds the *a-priori* determined number of latent variables that are most important to describe the data, but in a maximum likelihood sense (Bartholomew 2011 p. 181). But, where PCA assumes that the latent variables are orthogonal to each other, the factor analytic model can instead be adapted to all kinds of rotations (e.g. oblique, Bartholomew 2011 p. 70). In general, the factor analytic model is rotational invariant (Bartholomew 2011 p. 10), so that the solution can be rotated in any way without affecting the final fit. The most striking difference between PCA and FA is that PCA treats the row scores as fixed and FA as standard normally distributed random effects (Bartholomew 2011 p. 8).

FA is rarely applied in ecology (Kent & Ballard 1988; Kent 2006; Von Wehrden *et al.* 2009) since observations of ecological communities are not often normally distributed. A straightforward example is that of a count of individuals, which first and foremost counts cannot be negative. In such a case, unlike for normally distributed responses, patterns in the data are not well described by the sample covariance matrix (as in PCA), since it is a particularly naive estimator for the maximum likelihood solution of the covariance matrix for a latent variable model with a non-normal response distribution, so that generally more complex expressions are required in combination with iterative algorithms.

Popular unconstrained ordination methods in ecology

There are two different approaches in ordination to accommodate non-normal data and non-normal response models. Either: 1) a different distance measure is assumed to accommodate differences in data types, or 2) a different response distribution aside from normality is assumed. Roughly, these two groups can be considered as the most distinct branches in ordination methods to date, and as such we will examine each in more detail.

Distance-based ordination forms a separate group of methods within ordination that instead of directly working on the raw data, are applied to a matrix of distances or dissimilarities between the rows of a multivariate dataset. Use of the Euclidean distance measure, as in PCA, is discouraged in ecology for various reasons, relating to flaws of PCA for the analysis of ecological communities: e.g. the horseshoe effect. The horseshoe effect concerns the tendency of PCA to exhibit non-linear distortions, which is considered a mathematical artefact in ecology, and which is not due to any real ecological process that underlays the data generation (Swan 1970). As such, PCA is not recommended for the analysis of ecological communities

(Gauch & Whittaker 1972; Beals 1973; Kessell & Whittaker 1976; Nichols 1977; Rydgren 1996).

To improve on this, distance-based ordination methods instead rely on a $n \times n$ symmetric matrix of a type of *a-priori* chosen distances or (dis)similarities between rows, so that the information on the column identities is condensed (i.e. hidden, and cannot be retrieved in a meaningful way post-hoc). Retrieving a successful ordering of row coordinates with distance-based ordination methods thus largely depends on selecting a distance measure that appropriately accommodates the properties of a dataset. A range of studies have summarized the properties of dissimilarity measures, and for which data type they should be applied (Gauch & Whittaker 1972; Faith *et al.* 1987; Legendre & Gallagher 2001; Podani & Miklós 2002; Greenacre 2017; Roberts 2017). Faith *et al.* (1987) and Legendre & Gallagher (2001) studied the performance of some distance measures for ecological applications, and determined that various frequently used dissimilarity measures performed poorly.

PCoA performs an eigendecomposition of the matrix of dissimilarities, so that it is equivalent (up to the constant in the calculation for the sample covariance matrix) to PCA for an Euclidean distance (Mardia *et al.* 1980 p. 405). A dissimilarity measure is considered metric when a matrix is symmetric, has a diagonal of zeros, and where all off-diagonal entries are positive (Mardia *et al.* 1980 p. 395). The use of non-metric measures in PCoA can cause negative eigenvalues, which is considered one of the largest issues of the method (Cailliez 1983; Legendre & Legendre 2012 p. 501).

Note, that few modern advances have been made for distance-based ordination methods, such as t-distributed stochastic neighbour embedding (Van der Maaten & Hinton 2008). However, such methods have not yet been frequently applied in community ecology, and as such we do not discuss them here (but see e.g. Roberts 2020).

Non-metric multidimensional scaling

Arguably the most popular and well-known ordination method in ecology is NMDS. In contrast to PCoA, NMDS has been shown resistant to difficulties with mathematical artefacts (Minchin 1987), though see Ruokolainen & Salo (2006).

Unlike eigenvector methods, NMDS is an iterative algorithm that, just as eigenvector based methods, attempts to best project distances or dissimilarities between rows in the data, into a lower dimensional space where the distance between coordinates is Euclidean. However, NMDS has the unique property that the assumption of Euclidean distances in the ordination space can be adjusted (though this is rarely done to our knowledge). Minchin (1987) showed that NMDS performs especially well for ecological datasets, better than some other popular ordination methods (e.g. CA and DCA as discussed below).

At its core, the NMDS algorithm performs an (ordered by increasing values) isotonic regression of the original data dissimilarities on the dissimilarity matrix of row coordinates in the ordination space. Note that NMDS is prone to getting stuck in a suboptimal solution, so that it requires multiple fittings to ensure an optimal solution. Although the philosophy of NMDS is unlike other ordination methods, its properties are surprisingly similar to that of FA, since the dimension of the ordination space is also chosen *a-priori* to fitting in NMDS. Then, based on the chosen number of dimensions, NMDS finds the solution that minimizes the (scaled) squared residual of the rank order of the (dis)similarities in the data, and the rank order of rows in the ordination dimensional space. This objective function in NMDS is named "stress" (Kruskal 1964a,b). As such, the solution of NMDS does not directly depend on the original matrix of row dissimilarities, but only on the rank order of the rows, which is origin of the method its robustness to issues with mathematical artefacts.

In general, little research has been done on the statistical properties of NMDS. Brady (1985) attempted to do so by specifying different models, including:

$$\mathrm{d}_1(\boldsymbol{Y}) = \mathrm{f}\{\mathrm{d}_2(\boldsymbol{Z})\} + \boldsymbol{\epsilon}, \tag{5}$$

where $d_1(\cdot)$ and $d_2(\cdot)$ are distance functions that are not required to be the same, $f(\cdot)$ is a non-parametric monotonic function, $\boldsymbol{Z}$ a matrix holding vectors row scores $\boldsymbol{z}_i$, and $\boldsymbol{\epsilon}_i$ an normally distributed error term. Clearly, similarly to other ordination methods, NMDS finds the latent variables that best fit the data (albeit "fit" is measured in a different manner). This model serves to provide an impression of how NMDS relates to the other ordination methods described above. For further details on e.g. the development of a maximum likelihood estimator, and consistency of the NMDS estimator, see Brady (1985). Note, that here too the ordination is invariant to rotations, since for any two sets of row coordinates and with an orthogonal matrix $\boldsymbol{R}$ we can write $d_1(\boldsymbol{R}\boldsymbol{z}_1, \boldsymbol{R}\boldsymbol{z}_2) = d_1(\boldsymbol{z}_1, \boldsymbol{z}_2)$, so that the distances in the ordination do not change with changes in rotation or flipping of the axes.

Ruokolainen & Salo (2006) concluded that NMDS did not outperform eigenvector based methods, but instead drew the more nuanced conclusion that both groups of methods have their place for the analysis of community ecological data. This is similar to the "multiple parallel ordinations" procedure described by van Son & Halvorsen (2014).

Correspondence analysis

In this article, we use CA to connect the worlds of distance-based and model-based ordination. ter Braak & Barendregt (1986) and ter Braak (1985) showed that CA can be related to model-based ordination as a

type of latent variable model, so that rather than changing the distance measure to accommodate different data types as in PCoA and NMDS, we now change the response distribution for the latent variable model instead. Although, it is often noted that CA can be considered as a distance-based method too, namely by retrieving the solution by first calculating a matrix of χ^2 distances, and subjecting that to a singular value decomposition (Legendre & Legendre 2012 pp. 466–467), so that CA has a foot in both the distance-based and model-based ordination worlds.

CA was developed independently by a range of authors (Legendre & Legendre 2012 p. 464), most recently under the synonym "reciprocal averaging" (Hill 1973), for the analysis of contingency tables. Due to its simplicity and straightforward connection, below in equation (6a) we present the solution for CA that uses a weighted singular value decomposition of the data, where rows and columns are standardized by their sums, as presented by ter Braak (1985).

$$\text{diag}(\sum_{j=1}^{p} y_{ij})^{-0.5} \boldsymbol{Y} \text{diag}(\sum_{i=1}^{n} y_{ij})^{-0.5} = \boldsymbol{U} \text{diag}(\boldsymbol{\lambda}^{0.5}) \boldsymbol{V}^\top \tag{6a}$$

$$\boldsymbol{Y} = \text{diag}(\sum_{j=1}^{p} y_{ij})^{0.5} \boldsymbol{U} \text{diag}(\boldsymbol{\lambda}^{0.5}) \boldsymbol{V}^\top \text{diag}(\sum_{i=1}^{n} y_{ij})^{0.5}. \tag{6b}$$

Equation (6b) additionally serves to demonstrate how CA relates to the unstandardised data. The first eigenvector from this calculation represents an ordination axis where all loadings and the eigenvalue are one, i.e. a type of row specific intercept as in other latent variables models (see e.g. Hui 2016). This equally serves to demonstrate that the eigenvalues for CA are all in the range $\{0, 1\}$. The loadings of CA are $\boldsymbol{\gamma}_j = \text{diag}(\boldsymbol{\lambda}^{0.5}) \boldsymbol{v}_j (\sum_{i=1}^{n} y_{ij})^{-0.5}$ and row scores $\boldsymbol{z}_i = \boldsymbol{u}_i (\sum_{j=1}^{p} y_{ij})^{-0.5}$.

ter Braak & Barendregt (1986) derived conditions under which CA successfully approximates the maximum likelihood solution for a latent variable model with Poisson (with log-link function), Bernoulli (with logit-link function), or gamma (with inverse-link function) distributed responses, by repeated weighted averaging, and using unimodal responses. In ecology, the unimodal response model was considered more realistic than the linear model response model, so that CA was a more appealing ordination method to apply for ecologists than PCA (Hill 1973; Gauch 1982; ter Braak 2014). From that perspective, choosing between PCA or CA for the ordination of ecological data does not only indicate a change in response distribution, but also a change in model structure and complexity.

A few years after the introduction of CA in ecology by Hill (1973), ecologists had noticed flaws in CA some of which were similar to that of PCA (Gauch *et al.* 1977; Hill & Gauch 1980). In particular, CA can exhibit non-linear distortions which are not considered to be due to any real ecological patterns in the

data, which ecologists refer to as the arch effect (Legendre & Legendre 2012 pp. 482–483). Additionally, CA also suffers a phenomena referred as the edge effect, which is caused by the erroneous clustering of column loadings closer together at the ends of the latent variables (also see Hill & Gauch 1980 Fig. 2), as a result of partially unobserved niches for species (Hill & Gauch 1980).

Consequently, Detrended Correspondence Analysis (DCA, Hill & Gauch 1980) was developed as an attempt to to improve on these issues with the arch effect and the edge effect in CA. DCA detrends and performs non-linear rescaling of the axes to improve these issues. The non-linear rescaling often successfully manages to eliminate the arch effect (J. Oksanen, pers. comm., November 22nd 2018). Unfortunately, the detrending and rescaling in DCA lacks any clear theoretical or statistical basis, and the solution is sensitive to choice in the number of segments (Jackson & Somers 1991). An alternative method for removal of the arch effect is detrending by polynomials, although this process was abandoned quickly (Knox 1989).

Since its development, DCA has been controversial at best (Wartenberg *et al.* 1987; Peet *et al.* 1988; Knox 1989; Jackson & Somers 1991). DCA introduces yet another mathematical artefact referred to as the tongue effect (Minchin 1987), so that it is also not the most suited method for the dimension reduction of data on ecological communities. Removal of the arch effect can even lead to removal of real patterns in the data (Pielou 1984). Minchin (1987) showed that PCA and (D)CA perform poorly when faced with various degrees of non-linearity in the latent variable model, resulting in the arch and tongue effects. Various distance-based methods do not exhibit such issues, so that Minchin (1987) advocated for the use of NMDS instead.

Model-based ordination with unconstrained latent variables

More recently, Warton *et al.* (2012) and Warton & Hui (2017) studied various solutions of distance-based ordination methods, and concluded that those methods have the tendency to confound effects on the mean of a distribution and its variance, so that they should be applied to ecological data only with great care. In turn, Warton *et al.* (2015b) advocated for the use of more explicit statistical models in the analysis of data on ecological communities.

We have previously discussed one ordination method that poses an explicit statistical model, namely FA. FA is that ordination method with an explicit statistical method. Warton *et al.* (2015b) advocated for the use of more explicit statistical models in the analysis of data on ecological communities. However, for non-normally distributed responses the assumption of normally distributed residuals in FA is often not even approximately satisfied, and so the response distribution needs to be changed. This subsequently leads to a generalization of FA into Generalized Linear Latent Variable Models or GLLVMs (Skrondal & Rabe-Hesketh

2004), which allow the user to flexibly choose a response distribution and accommodate all common data types in ecology. GLLVMs are flexible in the model posed, so that it can be linear in form as in PCA and PCoA (Hui *et al.* 2015), and non-linear models for the latent variables are possible (see our discussion later on and in van der Veen *et al.* 2021b). In ecology, the default GLLVM assumed is:

$$\eta_{ij} = \beta_{0j} + \boldsymbol{z}_i^\top \boldsymbol{\gamma}_j, \tag{7}$$

where we recall that β_{0j} is an intercept for column j, $\boldsymbol{z}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ are normally distributed scores or latent variables for row i, and $\boldsymbol{\gamma}_j$ are the loadings for column j. When fitting GLLVMs, we have to integrate over the unobserved latent variables, so that the likelihood is given as:

$$\mathcal{L}(\Theta) = \sum_{i=1}^{n} \log\Bigg\{ \int_{-\infty}^{\infty} \prod_{j=1}^{p} f\Big(y_{ij}|\boldsymbol{z}_i, \Theta\Big) h\Big(\boldsymbol{z}_i\Big) d\boldsymbol{z}_i \Bigg\}, \tag{8}$$

where $f(y_{ij}|\boldsymbol{z}_i, \Theta)$ is the distribution of the responses given the latent variables, with mean $g(\eta_{ij})$ for a known link-function $g(\cdot)$ such as the log-link for Poisson responses or the logit-link function for binary responses, and where the vector Θ includes all parameters in the model, including any nuisance parameters as part of the distribution (e.g., the variance parameters $\boldsymbol{\sigma}_j^2$ in the Gaussian distribution).

Treating the latent variables as fixed effects instead of as random effects in a GLLVM, results in a solution that is similar to that of classical ordination method. We can demonstrate this by first assuming that the normal distribution for the latent variables has a mean with some variance $\boldsymbol{z}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. If we now assume that the covariance matrix equals zero, the latent variables simplify to means of the normal distribution, and equation (8) simplifies to the likelihood of any other multivariate GLM, but instead of including predictors provided by a researcher, they are estimated by maximizing the likelihood.

Though it is more straightforward to make such a comparison for model-based ordination and eigenvector based methods such as PCA, CA, model-based ordination are compared to distance-based ordinations on a regular basis and are expected to provide similar solutions by researchers (Hui *et al.* 2015; Popovic *et al.* 2019; Jupke & Schäfer 2020; Roberts 2020). Alternatively, GLLVMs are understood as modelling the residual covariance of the data on the linear predictor scale (Warton *et al.* 2015a). This is straightforward to see, as the combination of equation (7), and the standard normality assumption of the row scores, implies the residual variance $\boldsymbol{\gamma}_j^\top \boldsymbol{\gamma}_j$ for column j, and similarly for the covariance between two column, so that the size of the residual covariance matrix is $p \times p$, with the element for species (j, k) given as $\boldsymbol{\gamma}_j^\top \boldsymbol{\gamma}_k$ (Hui *et al.* 2015). The residual covariance matrix includes the dissimilarity between the columns, which ecologically is interpreted as associations (due to e.g., interacting species Ovaskainen *et al.* 2017). If instead a researcher

is interested in the predicted dissimilarity of rows, as in distance-based ordination, transposing the data (so that the rows become the columns and vice versa) and fitting the same model, results in a $n \times n$ residual covariance matrix of row dissimilarity. This is also the orientation used in e.g. the example of Gauch (1982) (pp. 138), where they used PCA to determine coordinates of rows, rather than coordinates of columns. Note however, that the resulting assumption of independent species is not appropriate.

Popular constrained ordination methods in ecology

So far, we have discussed ordinations that are constructed from information in the $n \times p$ response data matrix. Often however, measurements on the rows of the data matrix are additionally available in a separate matrix of predictors. There are multiple possibilities to include predictors in a multivariate analysis, which we discuss in this section.

Predictors can be directly included in multivariate GLM as in Wang *et al.* (2012):

$$\eta_{ij} = \beta_{0j} + \boldsymbol{x}_i^\top \boldsymbol{\beta}_j, \tag{9}$$

where $\boldsymbol{x}_i$ are the $k = 1 \dots K$ predictors for row i, and $\boldsymbol{\beta}_j$ are the corresponding slopes for column j of the responses. A multivariate GLM is a stack of independent univariate GLMs, so that the fixed effects parameters can be estimated by fitting models to each column of the data separately. Using GLLVMs it is possible to include an unconstrained ordination after including the predictors, as to perform an ordination on the residuals (i.e. left-over information, Carleton 1984) of the model:

$$\eta_{ij} = \beta_{0j} + \boldsymbol{x}_i^\top \boldsymbol{\beta}_j + \boldsymbol{z}_i^\top \boldsymbol{\gamma}_j, \tag{10}$$

where the last term on the right hand side is an ordination, which is $\boldsymbol{z}_i^\top \boldsymbol{\gamma}_j$, and has the same set up as in equation (7). However, here the ordination is conditional on the fixed effects, and for that reason such an ordination is referred to as a "residual ordination", since the latent variables still consist of residual information as in unconstrained ordination. However, unlike in an unconstrained ordination, a residual ordination excludes certain patterns from that unconstrained ordination, which are specified using the predictors. Typically, a residual ordination is used when the effect of a predictor needs to be accounted for, but is not of direct interest.

An alternative interpretation of residual ordination follows from the perspective that, after accounting for the effects of the predictors, there might be residual correlation left to explain between the columns of the

data. Depending on the research question at hand, these predictors may or may not be of interest, so that the residual ordination might be of more importance to answering a research question than the predictors (or vice-versa).

Constraining latent variables

ter Braak & Prentice (1988) unified unconstrained and constrained ordination in a single framework for ordination. Unlike in residual ordination, we now perform our inference on the latent variables, so that the latent variables can be considered as (partially) observed. Constrained ordination is so named, as it places constraints on the latent variables relative to latent variables made up from residual information alone, by using the measured predictor variables. For example, in constrained ordination we could assume that the latent variables are represented by the parallel change in various predictors simultaneously (Halvorsen 2012). Constrained ordination is the most popular way of including predictors in an ordination, at least in community ecology, and is a special case of a multivariate GLM or GLLVM.

Unconstrained or residual ordination methods are especially useful for community ecologists to generate hypotheses when measurements of the environment at sites are not available or of interest, whereas constrained ordination methods can serve also to test hypotheses on species-environment relationships (Økland 1996). Similar to unconstrained ordination, constrained ordination can be performed conditional on a set of predictors, a method that is referred to as partial constrained ordination (ter Braak 1988).

For constrained ordination methods we will assume the following model:

$$g\{\mathrm{E}(y_{ij}|\boldsymbol{x}_{lv,i},\boldsymbol{\epsilon}_i)\} = \beta_{0j} + \boldsymbol{z}_i^\top\boldsymbol{\gamma}_j; \qquad \text{where } \boldsymbol{z}_i^\top = \boldsymbol{x}_{lv,i}^\top\boldsymbol{B}, \tag{11}$$

where $\boldsymbol{x}_{lv,i}$ is a matrix of predictors for the ordination, and where $\boldsymbol{B}$ is a $d \times K$ matrix of predictor slopes common to all columns and rows of the data. This model can be fitted in e.g., the `VGAM` R-package (Yee & Hastie 2003), the `gllvm` R-package (Niku *et al.* 2017a), or using the `vegan` R-package (Oksanen *et al.* 2020). The ordination resulting from equation (11) can be visualized using a triplot (ter Braak & Verdonschot 1995), where arrows are drawn using the estimates of $\boldsymbol{B}$ (so that three quantities: rows, column and predictors are represented, hence a "triplot"), to represent the correlation of each predictor with the latent variable. The location of row scores and column loadings can then be interpreted in relation to those arrows.

Constrained ordination is a type of hierarchical regression where the latent variables are assumed to be (weighted) linear combinations of measured predictors. Indeed, Each of the classical ordination methods mentioned previously also has a constrained variant. For PCA this is Redundancy Analysis (RDA, Rao 1964), for NMDS it is constrained NMDS (Heiser & Meulman 1983), for (D)CA it is (Detrended) Canonical

Correspondence Analysis ((D)CCA, ter Braak 1986), and for model-based ordination it is reduced rank regression (Anderson 1951; ter Braak & Looman 1994; Yee 2015; van der Veen *et al.* 2021a). Constrained ordination can be considered as attempting to summarize information in two matrices at the same time (the responses and the predictors).

In a multivariate GLM as in equation (9), the rank of the matrix of predictor slopes is equal to the maximal number of linearly independent columns i.e. $\min(p, K)$ so that the number of parameters is $p+pK$. Reduced rank regression serves to reduce the number of parameters by imposing a rank constraint on the matrix of predictor slopes, with the additional benefit of being an ordination (ter Braak & Prentice 1988; van der Veen *et al.* 2021a). From equation (11) we can set $\boldsymbol{\beta}_{0j} = \boldsymbol{B}\boldsymbol{\gamma}_j$ to again retrieve a multivariate GLM as in equation (9). Since $\boldsymbol{z}_i$ can be interpreted as a latent variable, this leads to a similar procedure for selecting the rank constraint as in model-based unconstrained ordination. When $d < K$, the number of parameters is reduced compared to a multivariate GLM, and the dimension of both the response dataset and the matrix of predictors is reduced.

Classical constrained ordination methods such as RDA and CCA do not reduce dimension of the matrix of predictors, but instead relate predictors to the responses, and consecutively perform a maximum variance rotation. As such, methods such as RDA and CCA assume that $d = \min(K, p)$, i.e. the number of ordination axes is at maximum equal to the number of predictors or the number of column, whichever is least, so that dimension-reduction is performed post-hoc based on criteria for variances related to the latent variables as in unconstrained ordination. We now review each of these constrained ordination approaches in more detail.

Redundancy analysis

Redundancy analysis is the equivalent of multivariate regression, but adopts a post-hoc rotation to maximum variance for the fitted values. It was developed by Rao (1964), but termed by van den Wollenberg (1977). It is generally considered as the equivalent of PCA, but for constrained ordination (ter Braak & Šmilauer 2015). Similarly to in PCA, RDA has only limited applicability to ecological data, due to its linear response model and due to the assumption of multivariate normality (McCune *et al.* 2002).

In RDA, the data is first regressed against the predictors using multivariate regression, after which the matrix is subjected to a singular value decomposition of rank K, in order to retrieve dimensions that satisfy a maximum variance rotation, and to retrieve the row scores and column loadings for the ordination space. As such, we can formulate RDA as follows, for a column-centred matrix of predictors $\bar{\boldsymbol{X}}$ and a column-centred matrix of observations $\bar{\boldsymbol{Y}}$:

$$\bar{\boldsymbol{X}}\boldsymbol{\beta} = \boldsymbol{U}\text{diag}(\boldsymbol{\lambda}^{0.5})\boldsymbol{V}^\top; \qquad \text{where } \boldsymbol{\beta} = (\bar{\boldsymbol{X}}^\top\bar{\boldsymbol{X}})^{-1}\bar{\boldsymbol{X}}^\top\bar{\boldsymbol{Y}}(n-1)^{-1}, \tag{12}$$

In order to relate RDA to the model in equation (11), we now set $\boldsymbol{\beta} = \boldsymbol{BV}$ as in equation (11) (for $\boldsymbol{V} = \boldsymbol{\Gamma}$ with $\boldsymbol{\gamma}_j$ on the rows of that matrix) and note that, as for the latent variable models above, this solution is invariant to changes in rotation or scale, and that the maximum rank of $\boldsymbol{\beta}$ is K. Since we can set $\boldsymbol{BRR}^{-1}\boldsymbol{V}$ for any $K \times K$ non-singular $\boldsymbol{R}$, we see that $\boldsymbol{XRR}^{-1}\boldsymbol{BV} = \boldsymbol{U}\text{diag}(\boldsymbol{\lambda}^{0.5})\boldsymbol{V}^\top$ for $\boldsymbol{U} = \boldsymbol{XR}$, $\text{diag}(\boldsymbol{\lambda}^{0.5}) = \boldsymbol{R}^{-1}\boldsymbol{B}$, so that RDA is the equivalent of reduced rank regression for normally distributed responses and for a maximum variance rotation. ter Braak & Prentice (1988) (pp. 246) notes that for two ordination axes, RDA includes $p + d(K + p)$ parameters, which is exactly the case here, since there are p intercepts, $K \times d$ slopes for the predictors, and $p \times d$ column loadings: $p + K \times d + p \times d = p + d(K + p)$.

Similar to PCA, dimension-reduction is performed post-hoc in RDA, whereas in reduced rank regression the latent variables that provide the best fit are estimated, given the pre-selected number of latent variables that a researcher is interested in. For the latter case, the number of latent variables that optimally represent a dataset can be determined using information criteria.

Also, analogous to distance-bassed unconstrained ordination, a constrained ordination can be performed by first calculating a matrix of dissimilarities of rows. Distance-based constrained ordination methods have been developed based on RDA including transformation-based or distance-based RDA (tb-RDA and db-RDA, Legendre & Anderson 1999; Legendre & Gallagher 2001; Anderson & Willis 2003). In transformation-based RDA, RDA is applied directly to a transformed version of the data (Legendre & Gallagher 2001). Unlike db-RDA, tb-RDA provides both row scores and column loadings. Other distance-based methods in general only determine row scores, and not column loadings. For example, in distance-based RDA, a PCoA is first performed, after which RDA is applied to the results of PCoA.

Constrained Non-Metric Multidimensional Scaling

NMDS can be extended to a constrained ordination method, similarly to RDA for PCA (Heiser & Meulman 1983; McCune *et al.* 2002 p. 137), although we note that software developments are lacking (ter Braak & Šmilauer 2015 p. 687).

It is relatively straightforward to develop an algorithm for constrained NMDS by optimizing with respect to a set of (reduced rank) predictor slopes instead of a set of unconstrained row scores, as in latent variable models (van der Veen *et al.* 2021a), for example as in algorithm 1 where vech$(\cdot)$ is a half-vectorizing operator, converting a symmetric matrix to a vector, by retrieving its lower triangular entries. Translating the constrained NMDS algorithm to a software implementation in the `R` programming language can be done using derivative-free optimization, since isotonic regression is by default provided in `R` (see appendix S1 for an implementation of constrained NMDS).

Naturally, constrained NMDS inherits similar issues with the use of distances between rows as for uncon-

Algorithm 1 constrained NMDS

1: Specify the matrix of dissimilarities $dY = \text{vech}\{\text{d}_1(\boldsymbol{Y})\}$
2: Specify the number of ordination axes d
3: initialise the matrix of d slopes for K predictors
4: Order dY increasingly
5: $\boldsymbol{Z} = \boldsymbol{X}_i^\top \boldsymbol{B}$
6: Let $dZ = \text{vech}\{\text{d}_2(\boldsymbol{Z})\}$
7: Order dZ by the ordering of dY
8: Isotonic regression: $dY = f(dZ) + \epsilon$
9: let iY be the locations on the isotonic fit for the data
10: let iX be the fitted values for dZ
11: $\underset{\boldsymbol{B}}{\text{argmin}}\, s = \sqrt{\sum (iY - iX)^2 / \sum dY^2}$

strained ordination (Warton *et al.* 2012; Warton & Hui 2017). Additionally, NMDS places the constraint that the rank order of dissimilarities of rows in lower dimensional space should correspond to that of the data. Thus, in a situation where the dissimilarity of rows due to the predictors is different from the dissimilarity of rows due to the response data, the constraint placed by the predictors most likely prevents constrained NMDS from reaching an optimal solution. Naturally, this is an issue for classical constrained ordination methods in general, when important drivers of patterns in the data are missing, so that a residual effect should be added instead (see van der Veen *et al.* 2021a, and the discussion on this subject below).

Canonical Correspondence Analysis

Canonical Correspondence Analysis is the most popular constrained ordination method to date in community ecology (Von Wehrden *et al.* 2009; Warton *et al.* 2012). It was developed by ter Braak (1986) as the constrained counterpart of CA (but also see ter Braak 1987), and has been described as "the counterpart of RDA" (ter Braak 2014). CCA can be calculated in a similar manner as RDA (Legendre & Legendre 2012 p. 664), but for a matrix of observations $\boldsymbol{Y}^*$ which is weighted by row and column sums as in CA, and a matrix $\boldsymbol{X}^*$ which is row standardized and column centred, then:

$$\text{diag}(\sum_{j=1}^{p} y_{ij}^{0.5})\boldsymbol{X}^*\boldsymbol{\beta} = \boldsymbol{U}\text{diag}(\boldsymbol{\lambda}^{0.5})\boldsymbol{V}^\top; \qquad \text{where } \boldsymbol{\beta} = (\boldsymbol{X}^{*\top}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*\top}\boldsymbol{Y}^*, \tag{13}$$

so that CCA has a similar connection to equation (11) as RDA.[1]

Due to its suitability for common data types in ecology such as counts and binary data, CCA has been widely applied in community ecology. Similarly to CA, CCA has a quadratic "face", so that it has been used in order to account for non-linear responses (ter Braak 1986; Johnson & Altman 1999; Yee 2004). More

[1] After completion of the chapter, it was brought to our attention that equation (13) is not precise. For a precise formulation see ter Braak and Verdonschot (1995).

importantly, Palmer (1993) concluded that CCA performs well, as the arch effect rarely crops up due to the constraints on the latent variables, so that CCA performs well also for noisy real world settings in ecology (Jupke & Schäfer 2020; but see McCune 1997). Johnson & Altman (1999) concluded that variability of the reduced rank slopes $\boldsymbol{B}$ estimated by CCA is large, and care should be taken in their interpretation. As such, the reduced rank predictor slopes are rarely used, and the current `R` implementation in the `vegan` `R`-package, uses sample correlations between the estimated latent variables and the predictors to draw arrows in a triplot (see also ter Braak 1986).

Model-based ordination with constrained latent variables

For models fitted to non-normally distributed responses, such as those that CCA approximately fits, reduced Rank regression exactly fits the model in equation (11), but connects the linear predictor η_{ij} to the mean of the response distribution using a non-linear link function $g(\cdot)$, such as the log-link for counts. As such, each of the aforementioned (classical) ordination methods can be understood as (either exactly or approximately) fitting a model-based constrained ordination, with slight differences in scaling, standardisation of data, or rotation of the solution.

Model-based constrained ordination methods have been available for latent variable models for decades. For example, Yee (2004) developed a more exact counterpart for CCA using Reduced Rank Vector Generalized Linear Models. Most recently, van der Veen *et al.* (2021a) developed a method for constrained ordination that can include both fixed effects and random effects, for when the latent variable is only partially observed, which we further address in the next section.

GLLVMs: a flexible framework for ordination

The GLLVM framework has the ability to account for difficulties encountered in the application of classical ordination methods over the last century, which includes difficulties with the use of distance measures, the double-zero problem, and the arch effect. In the GLLVM framework researchers have the possibility to include additional random-effects to account for some of these issues, and more generally to find the most appropriate model structure in a non-arbitrary way by adapting known statistical model-building techniques.

In general, explicitly assuming a statistical model allows the mean-variance relationships of ecological data to be appropriately accounted for (see e.g. Wang *et al.* 2012), and offers access to standard tools for regression, such as residual diagnostics for checking assumptions (Hartig 2021), tools for model-selection (Burnham & Anderson 2002), and a principle means by which to account for the statistical uncertainties of all parameter estimates, not available when using classical ordination methods.

Model-based ordination completely circumvents the specification of a distance measure for the data, but instead requires specifying a probability distribution for the response as in equation (8). The probability distribution is chosen as to accommodate the properties of a dataset, and the choice is validated by checking residual assumptions. The main difference with using a distance or dissimilarity measure is that, rather than transforming the data, the mean of the data is transformed instead using a link function. Here, it is important to note that directly transforming data can have negative side-effects on the results of an analysis (O'Hara & Kotze 2010; Warton & Hui 2011; Warton 2018), so that adhering to the properties of data using a probability distribution is more appropriate.

That is not to say that difficulties with classical ordination methods cannot arise at all in model-based ordination, but merely that the GLLVM framework has the tools to adjust when those difficulties do arise. For example, the problem with double zeros, i.e. the tendency of unconstrained ordination methods to treat all zeros in a similar fashion while the underlying process generating the zeros can differ, can still be an issue. Ecologically, zeros can be generated when two different processes generated data, which in regression is acknowledged by assuming a zero-inflated or hurdle models (say Niku *et al.* 2017b; or Lambert 1992).

Similarly, the arch effect can crop up in model-based unconstrained ordination (Hui *et al.* 2015). Legendre & Legendre (2012) writes that it is due to the non-linear species responses to the ordination axis, while earlier Hill (1973) writes: "It (the arch effect) arises because the second axis (canonical variate) of reciprocal averaging is constrained to be uncorrelated with the first axis, but is in no way constrained to be independent of it", and similar arguments exist for other ordination methods. Since model-based ordination methods are a relatively new development, little research is available on how to use model-based ordination as a tool to address the problems that plague classical ordination methods. One of the few references to potential issues with the non-linear distortions in the ordination, as the arch effect is occasionally referred to, is in Hui *et al.* (2015).

The assumption of normality for the latent variables in GLLVMs is an appealing choice, also in light of the arch effect. It improves on the deficit of (P)CA that ordination axes are only required to be linearly independent (Gauch 1982 p. 153). That is, assuming (standard) normality of the latent variables, as is the default in GLLVMs, results in orthogonality of the latent variables. Fortunately, assuming orthogonality is equivalent to assuming independence for standard normal random variables, since for the case of $d = 2$ with $\boldsymbol{z}_i = (z_{i,1}, z_{i,2}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ we see that $\mathrm{E}(z_{i,1}z_{i,2}) = \mathrm{E}(z_{i,1})\mathrm{E}(z_{i,2}) = 0$, a result which is naturally independent of the choice for the number of latent variables.

However, that is not to say that the linear and quadratic (or higher order polynomial terms) are independent as well. van der Veen *et al.* (2021b) showed that under the assumption of multivariate normality, quadratic and linear terms of all latent variables are independent, so that assuming standard normality can

be considered as a form of simultaneous estimation and detrending by polynomials (ter Braak & Prentice 1988), and a similar result holds for higher order polynomials. Unfortunately, this same result serves to show that unless accounted for using an explicit quadratic model, a linear ordination can visualize the quadratic term as a separate ordination axis, since the independence assumption can still be adhered to.

Though the prior assumption for the latent variables is to be standard normally distributed, no such assumption is made for the (predicted) conditional distribution of the latent variables given the data $\mathrm{p}(\boldsymbol{z}_i|\boldsymbol{y}_i)$. The prediction for the latent variables is commonly used for the visualization of an ordination with GLLVMs. For example, GLLVMs fitted with Variational Approximations approximate the conditional distribution of the random-effect with a (fully) parametrized version of the normal distribution (referred to as the variational distribution, Hui *et al.* 2017), and Laplace's method makes similar assumptions (Niku *et al.* 2017b), so that we can expect arch-like distortions to also crop up in model-based ordination.

Due to the non-linear nature of the arch effect, adjusting the model structure to include quadratic terms can accommodate such issues. Alternatively, since an ordination is conditional on other terms in the model, variation that leads to the arch effect in an ordination can be accounted for by including additional terms, such as random effects in the form of random intercepts (Jamil & ter Braak 2013; Hui *et al.* 2015).

Other random effects can be included in the model to accommodate other properties of the dataset under study. For example, van der Veen *et al.* (2021a) extended constrained ordination by including a random effects term, so that the latent variables can be modelled as partially (un)observed, compared to the assumption of fully observed latent variables in classical constrained ordination. This addresses the difficulty of classical constrained ordination, where patterns in the data can be misrepresented if few predictors are measured. Their model unifies the constrained and unconstrained ordination frameworks, as it always optimally represents the ordination (in an unconstrained sense), even when few predictors are measured. Additionally, as in any other mixed effects model, structured random effects can account for non-independence of residuals due to e.g., nested or spatially structured sampling designs. For example, in the `gllvm` `R`-package (Niku *et al.* 2017a), but also in the `boral` `R`-package (Hui 2016) or the glmmTMB `R`-package (Brooks *et al.* 2017), optional random effects can be included with various structures including an autoregressive or spatial auto-correlation structure. Additionally, the `HMSC` `R`-package (Tikhonov *et al.* 2020) offers many other tools for random effects modelling with latent variables.

Finally, adding random effects naturally increases the complexity of models. Historically, there is a considerable difference in the method for determining the complexity of a latent variable model as fitted with eigenvector methods including PCA, D(C)CA and PCoA, in contrast to latent variable models fitted with methods that require selecting the number of dimensions prior to fitting, including NMDS, FA, and GLLVMs. Choosing the correct complexity of the model is critical to ensuring correct interpretation of an

ordination, and reduce undue computational burden. In eigenvector methods, the number of dimensions for inference is chosen based on an arbitrary threshold for the magnitude of the eigenvalues. One way to determine this threshold is using a screeplot, which is a type of barplot with the explained variance on the y-axis, and the latent variables on the x-axis, which provides a straightforward overview of reducing variance with an increase in the number of latent variables (Cattell 1966).

In model-based ordination, the correct level of model complexity is determined by finding the number of latent variables and random effects that provide the best fit. The model structure is then generally determined using information criteria (Preacher *et al.* 2013), or through other measures such as cross-validation or regularization (Bhattacharya & Dunson 2011; see also Hui *et al.* 2018). As such, it is possible to determine the number of dimensions for an ordination in a less arbitrary manner compared to eigenvector based methods. Unfortunately, it is also this issue that provides model-based ordination methods a steeper learning curve, as it requires researchers to be familiar with a wider range of tools for regression.

Discussion

Here, we have provided an overview of popular classical ordination methods and newly developed alternatives based on explicit statistical models. Pearson (1901) developed the first ordination method 120 years ago. Since then, many different ordination methods have been developed, come into use, and have been retired in favour of other ordination methods. For example, polar ordination (also known as Bray-Curtis ordination, Bray & Curtis 1957) is an ordination method that was frequently applied in the previous century in favour of PCA, but which has now fully been retired (Kent & Ballard 1988). When Hill & Gauch (1980) developed DCA to tend to the arch and edge effect issues in CA, that method was quickly adopted by ecologists, after which many researchers started to favour NMDS when Minchin (1987) showed that method to perform better. Confusingly, PCA remains a popular method of ordination, despite the many deficiencies it exhibits when used for the ordination of data on ecological communities (Swan 1970; Kent & Ballard 1988; Von Wehrden *et al.* 2009).

Similarly, Warton & Hui (2017) and Warton *et al.* (2012) studied the properties of distance measures and concluded that distance-based ordination methods perform poorly, though some researchers continue to favour those methods over more modern developments (see Roberts 2020; but also Hoegh & Roberts 2020). Distance-based ordination methods approximate the solution of a latent variable model by first calculating distances between the rows of the data, and afterwards attempting to retrieve latent variables. As a consequence, distance-based methods cannot estimate coordinates for columns in an ordination, making full exploration of patterns in the data impossible. Additionally, relating distance-based methods to predictors

is difficult since software implementations for distance-based constrained ordination are lacking, so that one of the few available alternatives to distance-based constrained ordination for non-normal data is post-hoc use of Generalized Additive Models (Wood 2017) in a NMDS (e.g. as implemented in the `vegan` R-package, Oksanen *et al.* 2020).

All ordination methods, whether distance-based or not, are used by ecologists to explain an underlying latent process in few dimensions. As such, applications of ordination methods in ecology can be understood as attempts to fitting a latent variable model, leading to the conclusion that model-based ordination methods provide a suitable alternative to all classical ordination methods. Model-based ordination methods are fully statistical nature, so that the statistical model can be adapted to accommodate any researchers' wishes. For example, Roberts (2020) writes that model-based ordinations are (too) heavily parametrized, yet developments to decrease the information burden and assume that certain parameters are the same for the whole community are well under way (see e.g. van der Veen *et al.* 2021b). One of the main benefits of the GLLVM framework is that it always has an explicit statistical model, that can be readily adapted to accommodate the hypotheses of any study or the wishes of any researcher. For example, Ovaskainen *et al.* (2017) include phylogenetic relatedness in their model to account for correlation between the columns of a multivariate dataset, and Niku *et al.* (2021b) includes functional traits. Thorson *et al.* (2015) includes spatial coordinates to account for non-independence between the row observations, while Tobler *et al.* (2019) accommodates an additional process model for imperfect detection. Further developments could consider better accommodating sparse ecological datasets, multiple datasets as in ter Braak & Schaffers (2004), or could instead account for error in the measurement of predictors in model-based ordination with constrained latent variables (McCune 1997).

With the developments of computational frameworks for fitting hierarchical models such as Template Model Builder (Kristensen *et al.* 2016) or NIMBLE (de Valpine *et al.* 2017), it has become more straightforward for quantitatively minded researchers to develop their own ordination methods as part of the GLLVM framework. However, mature and easy-to-use software implementations for model-based ordination methods are still a ways down the road for applied ecologists. The recent implementation of model-based ordination in the `glmmTMB` R-package (Brooks *et al.* 2017) is a promising development in that regard. Furthermore, the `gllvm` R-package (Niku *et al.* 2021a) provides various tools to perform (residual) model-based (un)constrained ordination for ecologists. The `Boral` R-package or the `HMSC` R-package provide yet again different tools to explore species distributions using the same latent variable approach (Hui 2016; Tikhonov *et al.* 2020).

Jupke & Schäfer (2020) argue that both model-based ordination and classical ordination have a place in the statistical toolset of ecologists. We instead wish to suggest that model-based ordination methods will be the next generation of ordination methods for the dimension reduction of multivariate datasets in ecology.

Classical ordination methods exhibit serious flaws that are difficult to adjust for in an analysis or when drawing inference. Model-based ordination methods improve on classical ordination methods in many ways, especially when it comes to the flexibility for future extensions (Hui 2017; Tikhonov *et al.* 2020; Damgaard *et al.* 2020; van der Veen *et al.* 2021a; van der Veen *et al.* 2021b), and with respect to validating assumptions that they make (Hui *et al.* 2015; Warton *et al.* 2015b). It is this flexibility that empowers ecologists to use the model-based ordination framework in a way that overcomes the weaknesses of the classical framework, as well as going beyond that framework to explore new questions in community ecology. Finally, it is important to recognize, that classical ordination methods are well studied, so that there exists a large body of literature on their application. That body of literature is vital for the understanding and application of model-based ordination, and will continue to be relevant in light of the modern developments discussed in this article.

Acknowledgements

We thank David Warton for helpful comments on earlier drafts of the manuscript. B.V. was supported by a scholarship from the Research Council of Norway (grant number 272408/F40). F.K.C.H was supported by an Australia Research Council Discovery Fellowship (grant number DE200100435).

References

Anderson, T.W. (1951). Estimating Linear Restrictions on Regression Coefficients for Multivariate Normal Distributions. *The Annals of Mathematical Statistics*, **22**, 327–351.

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

Austin, M.P. (1985). Continuum Concept, Ordination Methods, and Niche Theory. *Annu. Rev. Ecol. Syst.*, **16**, 39–61.

Bartholomew. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach, 3rd Edition*, 3rd editionn. Wiley, Chichester, West Sussex.

Beals, E.W. (1973). Ordination: Mathematical Elegance and Ecological Naivete. *Journal of Ecology*, **61**, 23–35.

Bhattacharya, A. & Dunson, D.B. (2011). Sparse Bayesian infinite factor models. *Biometrika*, **98**, 291–306.

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

Brady, H.E. (1985). Statistical consistency and hypothesis testing for nonmetric multidimensional scaling. *Psychometrika*, **50**, 509–537.

Bray, J.R. & Curtis, J.T. (1957). An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs*, **27**, 325–349.

Brooks, M.E., Kristensen, K., van Benthem, K.J., Magnusson, A., Berg, C.W., Nielsen, A., Skaug, H.J., Maechler, M. & Bolker, B.M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, **9**, 378–400.

Burnham, K.P. & Anderson, D.R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edn. Springer-Verlag, New York.

Cailliez, F. (1983). The analytical solution of the additive constant problem. *Psychometrika*, **48**, 305–308.

Carleton, T.J. (1984). Residual Ordination Analysis: A Method for Exploring Vegetation-Environment Relationships. *Ecology*, **65**, 469–477.

Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, **1**, 245–276.

Damgaard, C., Hansen, R.R. & Hui, F.K.C. (2020). Model-based ordination of pin-point cover data: Effect of management on dry heathland. *Ecological Informatics*, **60**, 101155.

de Valpine, P., Turek, D., Paciorek, C.J., Anderson-Bergman, C., Lang, D.T. & Bodik, R. (2017). Programming With Models: Writing Statistical Algorithms for General Model Structures With NIMBLE. *Journal of Computational and Graphical Statistics*, **26**, 403–413.

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

Gabriel, K.R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, **58**, 453–467.

Gauch, H.G. (1982). *Multivariate Analysis in Community Ecology*. Cambridge University Press, Cambridge.

Gauch, H.G. & Whittaker, R.H. (1972). Comparison of Ordination Techniques. *Ecology*, **53**, 868–875.

Gauch, H.G., Whittaker, R.H. & Wentworth, T.R. (1977). A Comparative Study of Reciprocal Averaging and Other Ordination Techniques. *Journal of Ecology*, **65**, 157–174.

Goodall, D.W. (1954). Objective methods for the classification of vegetation. III. An essay in the use of factor analysis. *Aust. J. Bot.*, **2**, 304–324.

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

Greenacre, M. (2017). Ordination with any dissimilarity measure: A weighted Euclidean solution. *Ecology*, **98**, 2293–2300.

Halvorsen, R. (2012). A gradient analytic perspective on distribution modelling. *Sommerfeltia*, **35**, 1–165.

Hartig, F. (2021). *DHARMa: Residual diagnostics for hierarchical (multi-level / mixed) regression models.*

Hastie, T., Tibshirani, R. & Friedman, J. (2016). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, 2nd editionn. Springer, New York, NY.

Hawinkel, S., Kerckhof, F.-M., Bijnens, L. & Thas, O. (2019). A unified framework for unconstrained and constrained ordination of microbiome read count data. *PLOS ONE*, **14**, e0205474.

Heiser, W.J. & Meulman, J. (1983). Constrained Multidimensional Scaling, Including Confirmation. *Applied Psychological Measurement*, **7**, 381–404.

Hill, M.O. (1973). Reciprocal Averaging: An Eigenvector Method of Ordination. *Journal of Ecology*, **61**, 237–249.

Hill, M.O. & Gauch, H.G. (1980). Detrended correspondence analysis: An improved ordination technique. *Vegetatio*, **42**, 47–58.

Hirschfeld, H.O. (1935). A Connection between Correlation and Contingency. *Mathematical Proceedings of the Cambridge Philosophical Society*, **31**, 520–524.

Hoegh, A. & Roberts, D.W. (2020). Evaluating and presenting uncertainty in model-based unconstrained ordination. *Ecology and Evolution*, **10**, 59–69.

Hui, F.K.C. (2016). Boral – Bayesian Ordination and Regression Analysis of Multivariate Abundance Data in r. *Methods in Ecology and Evolution*, **7**, 744–750.

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

Hui, F.K.C., Tanaka, E. & Warton, D.I. (2018). Order selection and sparsity in latent variable models via the ordered factor LASSO. *Biometrics*, **74**, 1311–1319.

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

Hui, F.K.C., Warton, D.I., Ormerod, J.T., Haapaniemi, V. & Taskinen, S. (2017). Variational Approximations for Generalized Linear Latent Variable Models. *Journal of Computational and Graphical Statistics*, **26**, 35–43.

Jackson, D.A. & Somers, K.M. (1991). Putting Things in Order: The Ups and Downs of Detrended Correspondence Analysis. *The American Naturalist*, **137**, 704–712.

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

Johnson, K.W. & Altman, N.S. (1999). Canonical correspondence analysis as an approximation to Gaussian ordination. *Environmetrics*, **10**, 39–52.

Jong, J.-C. & Kotz, S. (1999). On a relation between principal components and regression analysis. *The American Statistician*, **53**, 349–351.

Jongman, R., ter Braak, C. & van Tongeren, O. (Eds.). (1995). *Data analysis in community and landscape ecology*. Cambridge university press, Cambridge.

Jupke, J.F. & Schäfer, R.B. (2020). Should ecologists prefer model- over distance-based multivariate methods? *Ecology and Evolution*, **10**, 2417–2435.

Kent, M. (2006). Numerical classification and ordination methods in biogeography. *Progress in Physical Geography*, **30**, 399–408.

Kent, M. & Ballard, J. (1988). Trends and problems in the application of classification and ordination methods in plant ecology. *Vegetatio*, **78**, 109–124.

Kessell, S.R. & Whittaker, R.H. (1976). Comparisons of three ordination techniques. *Plant Ecol*, **32**, 21–29.

Kidziński, Ł., Hui, F.K.C., Warton, D.I. & Hastie, T. (2021). Generalized Matrix Factorization. URL http://arxiv.org/abs/2010.02469 [accessed 1 September 2021]

Knox, R.G. (1989). Effects of detrending and rescaling on correspondence analysis: Solution stability and accuracy. *Vegetatio*, **83**, 129–136.

Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H. & Bell, B. (2016). TMB: Automatic Differentiation and Laplace Approximation. *J. Stat. Soft.*, **70**.

Kruskal, J.B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, **29**, 1–27.

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

Lambert, D. (1992). Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing. *Technometrics*, **34**, 1–14.

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

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

Legendre, P. & Legendre, L. (2012). *Numerical Ecology*. Elsevier.

Lynn, H.S. & McCulloch, C.E. (2000). Using Principal Component Analysis and Correspondence Analysis for Estimation in Latent Variable Models. *Journal of the American Statistical Association*, **95**, 561–572.

Lynn, H.S., McCulloch, C.E. & others. (1995). A critique of maximum likelihood ordination using principal components analysis.

Mardia, K.V., Kent, J.T. & Bibby, J.M. (1980). *Multivariate Analysis*, 1st editionn. Academic Press, London ; New York.

McCune, B. (1997). Influence of Noisy Environmental Data on Canonical Correspondence Analysis. *Ecology*, **78**, 2617–2623.

McCune, B., Grace, J.B. & Urban, D.L. (2002). *Analysis of Ecological Communities*. MjM Software Design.

Minchin, P.R. (1987). An evaluation of the relative robustness of techniques for ecological ordination. *Theory and models in vegetation science: Proceedings of Symposium, Uppsala, July 8–13, 1985* (eds I.C. Prentice & E. van der Maarel), pp. 89–107. Advances in vegetation science. Springer Netherlands, Dordrecht.

Nelder, J.A. & Wedderburn, R.W.M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society: Series A (General)*, **135**, 370–384.

Nichols, S. (1977). On the interpretation of principal components analysis in ecological contexts. *Vegetatio*, **34**, 191–197.

Niku, J., Brooks, W., Herliansyah, R., Hui, F.K.C., Taskinen, S., Warton, D.I. & van der Veen, B. (2021a). *Gllvm: Generalized linear latent variable models*.

Niku, J., Brooks, W., Herliansyah, R., Hui, F.K., Taskinen, S., Warton, D.I. & van der Veen, B. (2017a). Package "gllvm.". *R Project*, **326**.

Niku, J., Hui, F.K.C., Taskinen, S. & Warton, D.I. (2021b). Analyzing environmental-trait interactions in ecological communities with fourth-corner latent variable models. *Environmetrics*, **32**, e2683.

Niku, J., Hui, F.K.C., Taskinen, S. & Warton, D.I. (2019). Gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in r. *Methods in Ecology and Evolution*, **10**, 2173–2182.

Niku, J., Warton, D.I., Hui, F.K. & Taskinen, S. (2017b). Generalized linear latent variable models for multivariate count and biomass data in ecology. *Journal of Agricultural, Biological and Environmental Statistics*, **22**, 498–522.

O'Hara, R.B. & Kotze, D.J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, **1**, 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. & Wagner, H. (2020). *Vegan: Community ecology package*.

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

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

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **2**, 559–572.

Peet, R.K., Knox, R.G., Case, J.S. & Allen, R.B. (1988). Putting Things in Order: The Advantages of Detrended Correspondence Analysis. *The American Naturalist*, **131**, 924–934.

Pielou, E.C. (1984). *The Interpretation of Ecological Data: A Primer on Classification and Ordination.* John Wiley & Sons.

Podani, J. & Miklós, I. (2002). Resemblance Coefficients and the Horseshoe Effect in Principal Coordinates Analysis. *Ecology*, **83**, 3331–3343.

Popovic, G.C., Warton, D.I., Thomson, F.J., Hui, F.K.C. & Moles, A.T. (2019). Untangling direct species associations from indirect mediator species effects with graphical models. *Methods in Ecology and Evolution*, **10**, 1571–1583.

Preacher, K.J., Zhang, G., Kim, C. & Mels, G. (2013). Choosing the Optimal Number of Factors in Exploratory Factor Analysis: A Model Selection Perspective. *Multivariate Behavioral Research*, **48**, 28–56.

Rao, C.R. (1964). The Use and Interpretation of Principal Component Analysis in Applied Research. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, **26**, 329–358.

Roberts, D.W. (2020). Comparison of distance-based and model-based ordinations. *Ecology*, **101**, e02908.

Roberts, D.W. (2017). Distance, dissimilarity, and mean–variance ratios in ordination. *Methods in Ecology and Evolution*, **8**, 1398–1407.

Ruokolainen, L. & Salo, K. (2006). Differences in performance of four ordination methods on a complex vegetation dataset. *Annales Botanici Fennici*, **43**, 269–275.

Rydgren, K. (1996). Vegetation-environment relationships of old-growth spruce forest vegetation in Østmarka Nature Reserve, SE Norway, and comparison of three ordination methods. *Nordic Journal of Botany*, **16**, 421–439.

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

Spearman, C. (1904). General intelligence, objectively determined and measured. *The American Journal of Psychology*, **15**, 201–292.

Swan, J.M.A. (1970). An Examination of Some Ordination Problems By Use of Simulated Vegetational Data. *Ecology*, **51**, 89–102.

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

ter Braak, C.J.F. (1985). Correspondence Analysis of Incidence and Abundance Data: Properties in Terms of a Unimodal Response Model. *Biometrics*, **41**, 859–873.

ter Braak, C.J.F. (2014). History of canonical correspondence analysis. *Visualization and verbalization of data*, pp. 61–75. Chapman and Hall/CRC, London.

ter Braak, C.J.F. (1988). Partial canonical correspondence analysis. *Classification and related methods of data analysis: Proceedings of the first conference of the International Federation of Classification Societies (IFCS), Technical University of Aachen, FRG, 29 June-1 July 1987*, pp. 551–558. North-Holland.

ter Braak, C.J.F. (1987). The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio*, **69**, 69–77.

ter Braak, C.J.F. & Barendregt, L.G. (1986). Weighted averaging of species indicator values: Its efficiency in environmental calibration. *Mathematical Biosciences*, **78**, 57–72.

ter Braak, C.J.F. & Looman, C.W.N. (1994). Biplots in Reduced-Rank Regression. *Biometrical Journal*, **36**, 983–1003.

ter Braak, C.J.F. & Prentice, I.C. (1988). A Theory of Gradient Analysis. *Advances in Ecological Research* (eds M. Begon, A.H. Fitter, E.D. Ford & A. Macfadyen), pp. 271–317. Academic Press.

ter Braak, C.J.F. & Schaffers, A.P. (2004). Co-Correspondence Analysis: A New Ordination Method to Relate Two Community Compositions. *Ecology*, **85**, 834–846.

ter Braak, C.J.F. & Šmilauer, P. (2015). Topics in constrained and unconstrained ordination. *Plant Ecol*, **216**, 683–696.

ter Braak, C.J.F. & Verdonschot, P.F.M. (1995). Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquatic Science*, **57**, 255–289.

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

Tikhonov, G., Opedal, Ø.H., Abrego, N., Lehikoinen, A., de Jonge, M.M., Oksanen, J. & Ovaskainen, O. (2020). Joint species distribution modelling with the r-package Hmsc. *Methods in ecology and evolution*, **11**, 442–447.

Tobler, M.W., Kéry, M., Hui, F.K.C., Guillera-Arroita, G., Knaus, P. & Sattler, T. (2019). Joint species distribution models with species correlations and imperfect detection. *Ecology*, **100**, e02754.

van den Wollenberg, A.L. (1977). Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika*, **42**, 207–219.

Van der Maaten, L. & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, **9**.

Van der Maaten, L. & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, **9**.

van der Veen, B., Hui, F.K.C., Hovstad, K.A. & O'Hara, R.B. (2021a). Model-based ordination with constrained latent variables.

van der Veen, B., Hui, F.K.C., Hovstad, K.A., Solbu, E.B. & O'Hara, R.B. (2021b). Model-based ordination for species with unequal niche widths. *Methods in Ecology and Evolution*, **n/a**.

van Son, T.C. & Halvorsen, R. (2014). Multiple parallel ordinations: The importance of choice of ordination method and weighting of species abundance data. *Sommerfeltia*, **37**, 1–37.

Von Wehrden, H., Hanspach, J., Bruelheide, H. & Wesche, K. (2009). Pluralism and diversity: Trends in the use and application of ordination methods 1990-2007. *Journal of Vegetation Science*, **20**, 695–705.

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

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

Wartenberg, D., Ferson, S. & Rohlf, F.J. (1987). Putting Things in Order: A Critique of Detrended Correspondence Analysis. *The American Naturalist*, **129**, 434–448.

Warton, D.I. (2018). Why you cannot transform your way out of trouble for small counts. *Biometrics*, **74**, 362–368.

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

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

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

Warton, D.I. & Hui, F.K.C. (2017). The central role of mean-variance relationships in the analysis of multivariate abundance data: A response to Roberts (2017). *Methods in Ecology and Evolution*, **8**, 1408–1414.

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

Whittaker, R.H. (1967). Gradient Analysis of Vegetation. *Biological Reviews*, **42**, 207–264.

Wood, S.N. (2017). *Generalized Additive Models: An Introduction with R, Second Edition*. CRC Press.

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

Yee, T.W. (2015). *Vector Generalized Linear and Additive Models: With an Implementation in R*. Springer.

Yee, T.W. & Hadi, A.F. (2014). Row–column interaction models, with an R implementation. *Comput Stat*, **29**, 1427–1445.

Yee, T.W. & Hastie, T.J. (2003). Reduced-rank vector generalized linear models. *Statistical Modelling*, **3**, 15–41.

Økland, R.H. (1996). Are ordination and constrained ordination alternative or complementary strategies in general ecological studies? *Journal of Vegetation Science*, **7**, 289–292.

Next generation ordination with Generalized Linear Latent Variable Models

Bert van der Veen[123] Knut A. Hovstad[34] Francis K.C. Hui[5]
Robert B. O'Hara[23]

[1]Department of Landscape and Biodiversity, Norwegian Institute of Bioeconomy research, Trondheim, Norway
[2]Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway
[3]Centre of Biodiversity Dynamics, Norwegian University of Science and Technology, Trondheim, Norway
[4]The Norwegian Biodiversity Information Centre, Trondheim, Norway
[5]Research School of Finance, Actuarial Studies and Statistics, The Australian National University, Canberra, Australia

Appendix S1: Code for constrained NMDS

```
library(nloptr)
library(pracma)
library(vegan)


stress <- function(x, dY, d, X, type) {
    # x is a vector of parameter values dY is a symmetric matrix of
    # dissimilarities d is the number of axes X is a matrix of predictors


    if (!is.matrix(dY)) {
        dY <- as.matrix(dY)
```

```
}
# Step 1: Vectorize distance matrix
if (type == "global") {
    dY <- as.vector(dY[lower.tri(dY)])
} else if (type == "local") {
    dY <- t(dY)
    dY <- dY[col(dY) != row(dY)]
}

# add global local etc

# Step 2: d is the number of ordination axes

# Step 3: Organise matrix of coefficients
B <- matrix(x, ncol = d)

# Order distance matrix
ord <- order(dY)
dY <- dY[ord]

# Step 5: Calculate ordination axis
Z <- X %*% B

# Step 6: Calculate (euclidean) distances in the lower dimensional
# space
dZ <- dist(Z)
dZ <- as.matrix(dZ)

if (type == "global") {
    dZ <- as.vector(dZ[lower.tri(dZ)])
} else if (type == "local") {
    dZ <- t(dZ)
```

```
        dZ <- dZ[col(dZ) != row(dZ)]
    }

    # Step 7: Order dZ by the order of dY
    dZ <- dZ[ord]

    # Step 8: Isotonic regression. This is where the magic happens
    iReg <- isoreg(dY, dZ)

    # Step 9
    iY <- iReg$y

    # Step 10
    iX <- iReg$yf

    # Specify stress
    s <- (iY - iX)^2
    s <- sqrt(sum(s)/sum(iY^2))

    return(s)
}

doCNMDS <- function(Y, distance = "bray", d, X, type = "global", seed = NULL,
    n.init = 20, opt.control = list(maxeval = 1000, xtol_rel = 1e-08)) {
    X <- as.matrix(X)
    # Generate initial values
    fa <- factanal(Y, factors = d, scores = "regression")
    initMod <- lm(fa$scores ~ -1 + X)
    # jitter starting values a little
    dY = vegdist(Y, method = distance)

    resBest <- NULL
```

```
    if (n.init > 1)
        seed <- sample(1:10000, n.init)
    for (i in 1:n.init) {
        set.seed(seed[i])
        init <- c(coef(initMod) + matrix(rnorm(d * ncol(X), sd = 0.05),
            ncol = d))

        res <- nloptr::bobyqa(init, stress, dY = dY, d = d, X = X, type = type,
            control = opt.control)
        if (!is.null(resBest)) {
            if (res$value < resBest$value) {
                resBest <- res
            }
        } else {
            resBest <- res
        }
    }

    B <- matrix(resBest$par, ncol = d)
    row.names(B) <- colnames(X)
    colnames(B) <- paste("Ordination axis", 1:d, sep = "_")
    # finite differences approximation for variance of estimates hess <-
    # pracma::hessian(stress,res$par, dY = dY, d = d, X = X, type = type)
    # se <- sqrt(diag(solve(hess))) se <- matrix(se, ncol = d) colnames(se)
    # <- colnames(B) row.names(se) <- row.names(B) return(list(B = B, se =
    # se))
    return(list(B = B, stress = resBest$value, iter = resBest$iter, convergence = resBest$convergence))
}

# example
library(mvabund)
data(spider)
```

```r
y <- as.matrix(spider$abund)
X <- model.matrix(~-1 + ., data = data.frame(scale(spider$x)))
result <- doCNMDS(spider$abund, d = 2, X = X)
B <- result$B

LV <- X %*% B
plot(LV, type = "n")
text(LV)

# Do some funky scaling stuff
marg <- par("usr")
origin <- c(mean(marg[1:2]), mean(marg[3:4]))
Xlength <- sum(abs(marg[1:2]))/2
Ylength <- sum(abs(marg[3:4]))/2

B <- B/max(abs(B)) * min(Xlength, Ylength) * 0.8

# Make a plot
arrows(x0 = origin[1], y0 = origin[2], y1 = B[, 2] + origin[2], x1 = B[,
    1] + origin[1], col = "red")
text(x = (B[, 1] + origin[1]) * (1 + 0.2), y = (B[, 2] + origin[2]) * (1 +
    0.2), labels = colnames(X), col = "red")
```

3.2 Hierarchical Ordination, A unifying framework for drivers of community processes

Hierarchical Ordination, A unifying framework for drivers of community processes

R.B. O'Hara[12] B. van der Veen[123]

[1]Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

[2]Centre of Biodiversity Dynamics, Norwegian University of Science and Technology, Trondheim, Norway

[3]Department of Landscape and Biodiversity, Norwegian Institute of Bioeconomy research, Trondheim, Norway

Abstract

Ordination methods have been used by community ecologists to describe and explore the communities they sample by reducing this variation down to a small number of dimensions. More recently, Joint Species Distribution Models have been developed to model and predict the distributions of several species simultaneously. Contemporary models for the data for both of these problems are essentially the same, called Generalised Linear Latent Variable Models (GLLVMs). Based on this we suggest some avenues of cross-fertilisation between the two areas of research. We also describe some of the extensions to GLLVMs, and from this suggest the development of Hierarchical Ordination, as a way of efficiently modelling communities of species in space. **keywords**: Ordination, JSDM, GLLVM

Word count: 4039

Introduction

Community ecology is the study of groups of species. One important set of questions relate to how and why species are distributed relative to each other. For example, do particular species tend to occur together, and can this be explained by similar responses to the environment. These same questions can be asked across

different scales, e.g. by looking at samples from different Dutch sand dunes, or the complete distributions of different species across the globe. Statistical methods have been developed to look at this type of data for both of these scales, but their similarities have not been widely appreciated.

The methods we consider in this article assume that species are observed at a number of sites, so the data is a site by species matrix, with entries being a measure of abundance or incidence, e.g. whether a species was observed on that site, or the number of individuals observed. Historically, these data have been summarised by **ordination methods** (e.g. Gower 1966; ter Braak 1985), which look to reduce the variation between sites down to a small number (usually two) of dimensions, so that similar sites are closer together. This means that they can be plotted on an **ordination diagram**, which can be inspected visually. Of relevance here, a model-based approach to ordination has been developed, leading to Generalised Linear Latent Variable Models (Hui *et al.* 2015).

The analysis of communities has been approached by extending species distribution models (SDMs). SDMs were developed to look at the distributions of single species, but the perceived importance of species interactions (Kissling *et al.* 2012; Wisz *et al.* 2013) lead to the development of Joint Species Distribution Models (JSDMs: Pollock *et al.* 2014), which could incorporate several species. These model the response of each species to the environment, and then add a correlation matrix to allow for additional covariance between species. It was quickly realised that this matrix becomes unmanageable for many species, as it has so many parameters to estimate. Thus, extensions were developed to model the matrix as the sum of a smaller set of linear effects (e.g. Warton *et al.* 2015; Ovaskainen *et al.* 2017b). Because these models are based on a formal probabilistic model (which can be written as a likelihood, and thus fitted to data with flexible statistical methods), they have been extended in a variety of directions (see below).

The ecological differences between ordination and JSDMs largely revolve around the aims of the data collection and analysis, which has been a reflection of the types of ecologist using them. Ordination has typically been used by field ecologists wanting to describe the differences between communities they have sampled at different sites or different times, and possibly look at how these differences are correlated with environmental variables. Thus the focus has been on exploration of correlations between species and communities. In contrast, JSDMs were developed by and for macro-ecologists, wanting to look at the full distributions of species as extensions of SDMs, with the primary aim of estimating the effects of environmental covariates on the distribution of species, and thus being able to predict the current and future distributions of the species. The role of the residual correlation was to improve the predictions, rather than to be interpretable: problems with interpreting the correlations as interactions between species were pointed out early on (Pollock *et al.* 2014).

Thus, JSDMs tend to focus on a larger spatial scale, and have a bigger emphasis on the effects of covariates. As a practical matter, they are also generally based on presence/absence (but see Björk *et al.* 2018a for a counter-example), whereas ordination has been carried out on a wider range of data. Despite these differences, the underlying idea - to reduce covariation down to a smaller number of dimensions - is the same. Here we will point out the similarities between the two approaches, and use this to suggest that each can learn from the other. We then suggest a scheme to flexibly extend these methods without making them too unwieldy.

The Models

We will first develop a model that is a simple example of model-based ordination and JSDMs, and use this to explain the more complex models that have been developed.

In the simplest case, the data are a matrix, with rows being sites, and columns being species. The entries are observations of the species, which can take several forms, e.g. counts, percent cover, or (particularly for JSDMs) binary presence/absence. In addition to this, we can have covariates which are associated with the rows or columns, such as measures of the environment (temperature, habitat etc.) for each site, or traits (size, diet etc.) for each species.

From this data we can create a straightforward model, and then extend it. Each observation of species j ($j = 1, \ldots, p$) at sites i ($i = 1, \ldots, n$) is denoted y_{ij}. We call the full matrix of observations Y. We can model Y as an extension of a generalised linear model, with $g(E(y_{ij})) = \eta_{ij}$, where $g(\cdot)$ is a link function (e.g. logit or cloglog for presence/absence). On the link scale we model the expected value of each datum as

$$\eta_{ij} = \alpha_i + \phi_j + \varepsilon_{ij}, \tag{1}$$

where α_i is the site effect for site i, and ϕ_j is species effect of species j. ε_{ij} is an error term, which is modelled as being correlated between species, i.e.

$$\boldsymbol{\varepsilon}_i \sim MVN(\mathbf{0}, \Sigma), \tag{2}$$

where Σ is a $p \times p$ covariance matrix, with off-diagonal terms being the covariances. The number of parameters in Σ increases quadratically with the number of species, which makes the matrix difficult to estimate for large p, and the large number of parameters also makes the matrix difficult to interpret. The ordination/JSDM approach to handling this is to write the matrix as the sum of a product of row- and column- effects. Each

product defines one of L *latent dimensions*. The model is then

$$\eta_{ij} = \alpha_i + \phi_j + \sum_{l=1}^{L} z_{il}\gamma_{jl}. \tag{3}$$

In ordination, z_{il} is called a site score (or latent variable), and γ_{jl} is called a species loading. We can interpret z_{il} as a location along environmental gradient l, or informally that there are L unobserved covariates, with γ_{jl} being the regression parameter for the effect of covariate l on species j. The scale is arbitrary, so it is convenient to scale the site scores so that $\text{Var}(z_{il}) = 1$. Then, the covariance between species 1 and 2 becomes $\sum_{l=1}^{L} \gamma_{1l}\gamma_{2l}$. This model can be extended in many ways, as described below, by putting further models on α_i, ϕ_j, z_{il}, and γ_{jl}.

Adding Covariates

The model described above projects the correlation matrix down into L dimensions, where L is relatively small. But, as written, it does not include covariates. There are several ways to add them: for example, Ovaskainen *et al.* (2017b) outlined an approach where α_i and ϕ_j are made functions of species- and site-level covariates respectively, i.e. we could use a model such as

$$\eta_{ij} = \sum_{k=1}^{K} \beta_{jk}X_{ik} + \sum_{s=1}^{S} \theta_{js}W_{js} + \sum_{l=1}^{L} z_{il}^{\top}\gamma_{jl}, \tag{4}$$

where there are K site-level covariates (e.g. climate variables), and S species-level covariates (e.g. traits). The differences between some of the terms is, whether the terms are known (i.e. covariates), or whether they have to be estimated. Thus, the model incorporates responses of species to environmental conditions, as well as the effects of species' characteristics on their site response.

The regression coefficient matrices $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ can become large if there are many covariates, so some way to reduce this is desirable. This is simply a variable selection problem, for which there are several solutions available, through either selecting which variables are "in" and "out", or some form of regularisation (e.g. O'Hara & Sillanpää 2009 for some Bayesian methods; Tredennick *et al.* 2021 for some non-Bayesian approaches). Random effects can be added, for example β_{jk}, the effect of the k^{th} covariate on species j can be be modelled as to take phylogenetic correlation into account, i.e. so that species with a more recent common ancestor tend to have more similar values of β_{jk}. This approach was developed more fully by Ovaskainen *et al.* (2017b).

In classical ordination, covariates are included through **constrained ordination** (e.g. ter Braak 1986), where the site scores are forced to be linear functions of covariates, i.e.

$$z_{il} = \sum_{k=1}^{K} X_{ik}\psi_{kl}. \tag{5}$$

This implies that the latent variables are fully explained by the observed covariates. van der Veen *et al.* (2021a) extended this approach to allow a latent variable to be affected by covariates, plus additional (unmodelled) effects:

$$z_{il} = \sum_{k=1}^{K} X_{ik}\psi_{kl} + \varepsilon_{il}. \tag{6}$$

This assumes that the environmental gradient is affected by (or, at least, correlated with) measures of the environment. Random effects can also be added, in particular the gradient can be modelled spatially as a continuous field, leading to a type of **spatial factor analysis** (Thorson *et al.* 2015), where sites that are closer to each other tend to have more similar site effects. In essence, both fixed and random effects can be used to put structure on z_{il}. Another variation on this is to use more than one set of latent variables, with some responding to the environment, and others being unconstrained. For example, Björk *et al.* (2018a) modelled host-associated microbiota by including separate ordinations for the host species and the samples (within host species).

The **response** of a species to the gradient can also be modelled as being affected by the environment (e.g. Tikhonov *et al.* 2017; Perrin *et al.* 2021):

$$\gamma_{ijl} = \sum_{i=1}^{n}\sum_{k=1}^{K} X_{ik}\psi_{jkl}. \tag{7}$$

This allows the species effect to depend on the environment, so that the covariance between species j and h on site i is proportional to $\sum_{k=1}^{K} X_{ik}\psi_{jkl}\psi_{hkl}$, i.e. is a linear function of the covariate. So the correlation can change across environments, and potentially change sign.

Learning From Each Other

Within the development of the GLLVM framework, we can see that JSDMs and model-based ordination are equivalent mathematically. The biological differences stem from the data and the questions being asked.

Ordination has typically looked at a set of sites that have been sampled, and asks about the similarities between the communities on these sites. JSDMs, in contrast, try to model and predict the distributions of the species across their whole ranges, i.e. they operate at a larger spatial scale, and also are interested in the full distribution rather than a subset of sites. So for ordination, the relationship between species (or of sites) is in focus, whereas JSDMs are primarily intended to predict the distribution of each species. In contrast, JSDMs are primarily intended to predict the distribution of each species: the covariance between species is mainly of interest because it improves the prediction (Wilkinson *et al.* 2021).

The equivalence between ordination and JSDMs suggests that each area should be able to help the other. At the conceptual level, the questions being asked in the ordination world can also be asked by the SDM world. Most usefully, the methods and theory developed in one area can be transferred across and used in the other.

What can JSDMs learn from Ordination?

The focus of ordination has usually been on the whole community, rather than looking at individual species. Because of this, the typical summaries are visualisations, i.e. ordination plots, which condense information from sites and species into plots that can be interpreted. These can help with understanding co-occurrence patterns, and help to guide further modelling and interpretation (e.g. if several species cluster together). Using ordination plots into the analysis of JSDM should thus help with summarising and interpreting the correlations between species and sites.

The ecological interpretation of ordinations is also more advanced. If the ordination axes are interpreted as ecological gradients, then the site scores represent the locations of the species on the gradient, and the species loadings are the species' optima. Thus the ordination axes can be interpreted as part of the species's niche. Mathematically, the ordination model is a simplification of the species packing model (Jamil & Ter Braak 2013), where each species has an equal tolerance to the gradient. Adding a quadratic term in the ordination relaxes the equal tolerance assumption, so species can be generalists or specialists with respect to the gradient (van der Veen *et al.* 2021b). On top of this, a constrained ordination, can be used to efficiently model the niches of many species together. This will be useful when a large number of species are being considered, e.g. from meta-barcoding data, and particularly when properties of the whole community, rather than of each species, will be important.

What can Ordination learn from JSDMs?

Ordination can take advantage of the flexibility of the modelling framework, which has been developed more fully in the JSDM world. As described above, Ovaskainen *et al.* (2017b) developed a framework to incorporate a wide range of effects on both the species and sites, and spatial effects on the site effects can be added through spatial factor analysis (Thorson *et al.* 2015). The statistical modelling framework that has been used in the development of JSDMs is explicitly linear, which makes it straightforward to write down models with extra effects. Although fitting these models may not be easy, flexible Bayesian packages such as JAGS (Plummer 2021) and NIMBLE (de Valpine *et al.* 2017) can be used to specify these models in a straightforward language.

The JSDM approach also helps with incorporating better sampling models (e.g. Beissinger *et al.* 2016; Björk *et al.* 2018a; Tobler *et al.* 2019), so that having replicate information at a site, or other aspects of sampling design, can be properly incorporated into the model. Thus, for example, multiple traps or visits at a site can be treated as replicate samples from the same community: with binary data this can be used in an occupancy model (Tobler *et al.* 2019), but when the data estimate abundance (e.g. through counts), the repeated samples can be used to estimate the amount of sampling error.

Another area where ordination can follow methods developed in the JSDM world is the use of temporally explicit models. Because the approach is based on an explicit model, it can include a temporal autocorrelation (Ovaskainen *et al.* 2017a; e.g. Björk *et al.* 2018b). This is preferable to the approach that classical ordination takes, where the order of the years is ignored in the ordination, although it can be incorporated into the graphical presentation (e.g. Blanchette & Pearson 2013). This links the temporal changes more directly to models of community dynamics, as it is a multi-species Gompertz model where environmental effects on growth rates can be incorporated (e.g. Mutshinda *et al.* 2011).

Hierarchical Ordination; a unifying framework for drivers of community processes

The GLLVM framework is flexible enough to be developed in several directions. Site scores have already been modelled a functions of covariates, through a constrained GLLVM (van der Veen *et al.* 2021a), and also as a spatial field (allowing for spatial autocorrelation), in a spatial factor analysis (e.g. Thorson *et al.* 2015). If we transpose the data matrix, we can model the species effects, e.g. as traits. This can thus link ordination to trait-based analyses.

In the fixed effects, estimating the interactions between site and species covariates are the "fourth corner problem". This can be parameter heavy when there are several covariates (e.g. see Niku *et al.* 2021 in the context of GLLVMs). One possible approach is to extend the constrained ordination idea, so that both site scores and species loading are modelled further. This leads us to the idea of a hierarchical ordination, an extension of double constrained ordination (ter Braak *et al.* 2018). As before we have

$$\eta_{ij} = \alpha_i + \phi_j + \sum_{l=1}^{L} z_{il}\gamma_{jl}, \tag{8}$$

But now we extend the models for both z_{il} and γ_{jl}:

$$z_{il} = \sum_{k=1}^{K} X_{ik}\psi_{kl} + \varepsilon_{il}, \tag{9}$$

and

$$\gamma_{jl} = \sum_{s=1}^{S} W_{js}\omega_{sl} + \epsilon_{jl}. \tag{10}$$

Because both z_{il} and γ_{jl} can be written as sums of other terms, it is straightforward to model them both hierarchically, as we would for any other hierarchical model. For example, a spatial effect can be added to the site scores, similar to the spatial factor analysis idea, but trait and phylogenetic effects can also be added to the species loadings. The method developed by ter Braak *et al.* (2018) is similar, but they assume $z_{il} = \sum_{k=1}^{K} X_{ik}\psi_{kl}$ and $\gamma_{jl} = \sum_{s=1}^{S} W_{js}\omega_{sl}$, i.e. the site scores and species effects are fully determined by the covariates.

Expanding the model, and for clarity using only one latent variable and one species- and site- covariate, we get

$$\begin{aligned}\eta_{ij} &= \alpha_i + \phi_j + (\boldsymbol{\psi}^\top \boldsymbol{X}_i + \boldsymbol{\varepsilon}_i)^\top (\boldsymbol{W}_j\boldsymbol{\omega} + \boldsymbol{\epsilon}_j) \\ &= \alpha_i + \phi_j + \boldsymbol{X}_i^\top \boldsymbol{\psi}\boldsymbol{W}_j\boldsymbol{\omega} + \boldsymbol{\varepsilon}_i^\top \boldsymbol{W}_j\boldsymbol{\omega} + \boldsymbol{X}_i^\top \boldsymbol{\psi}\boldsymbol{\epsilon}_j + \boldsymbol{\varepsilon}_i^\top \boldsymbol{\epsilon}_j.\end{aligned} \tag{11}$$

This model acts as an efficient model for the fourth corner effect because it incorporates trait by environment interactions by relating them through the latent variables, rather than directly interacting with each other. Thus not all trait and species combination needs to be considered. Going through the ordination terms we have

- $\boldsymbol{X}_i^\top \boldsymbol{W}_j \boldsymbol{\psi\omega}$: the fourth-corner term, approximated in reduced rank,
- $\boldsymbol{\varepsilon}_i^\top \boldsymbol{W}_j \boldsymbol{\omega}$: a constrained ordination where the species loadings are constrained by traits alone,
- $\boldsymbol{X}_i^\top \boldsymbol{\psi\epsilon}_j$: a constrained ordination with site effects determined by environmental covariates,
- $\boldsymbol{\varepsilon}_i^\top \boldsymbol{\epsilon}_j$: a residual ordination.

Each of which would have a different place in the analysis of co-occurrence patterns in ecological communities. Of course, these terms can also include structured random effects.

Another way to look at this model is that it can model the change in associations between species with the environment (Tikhonov *et al.* 2017; Perrin *et al.* 2021). This comes from the fourth corner term, which models the interaction between traits and environment, so plays the same role as equation (7).

One advantage of having an explicit hierarchical ordination framework is that visualisation and interpretation can be based on current ordination methods, for example biplots can be drawn for both site and species effects. Thus the advantages of ordination as a way of summarising correlation, and the effects of covariates on the correlations, are retained with this model.

There may be problems where a single hierarchical ordination is not sufficient. For example, a species-specific covariate effect may be needed, leading to this model:

$$\eta_{ij} = \alpha_i + \phi_j + \sum_{k=1}^{K} X_{ik}\beta_{jk} + \sum_{l=1}^{L} z_{il}^\top \gamma_{jl}. \tag{12}$$

Potentially, more than one ordination could also be used. For example Björk *et al.* (2018a) analysed host-associated microbiota in a single model with ordinations at both the sample and host species levels. An equivalent model using the hierarchical ordination approach proposed here would use the same species loadings for both levels. Thus the framework is flexible, although there is a cost in adding extra latent levels, both computationally and in terms of interpretation. Using a single ordination, with effects on either species or sites, we simplify the model into manageable parts. The choice of whether to use one or two sets of ordination will depend on whether a single ordination is reasonable ecologically, and whether multiple ordinations are feasible computationally. Whilst the complexity of additional ordinations is attractive, the price is that more data will be needed to estimate the parameters of the model, and the fitting will be more difficult.

Discussion

We have shown that ordination and JSDMs have converged onto a similar set of models, GLLVMs, providing a common framework that is available to both groups of ecologists. The differences between the two approaches largely came from scale and the questions being asked, but the models are flexible enough to handle these. Having a single set of models should help to unify these different approaches to community ecology, and the link to temporal models should improve connections to ecological theory, efficiently integrating data into the frameworks that are being developed.

A model-based approach allows for a lot of flexibility, as we can see in our overview. One downside of this flexibility is the possibility that the model becomes too complex to be interpretable. With a large number of species and sites, it is easy to develop models that ask for all of this information to be used. Our suggestion of a hierarchical approach to ordination tries to reduce this complexity by making the parameters either site- or species- specific. Interactions between the two sets of parameters are made through the latent variables. This is one simplification of the model, so the response of a community to changes in the environment are measured by how the site scores change, and then how this affects species, through their loadings. This shifts focus from individual species or sites to the community, and the gradients.

It is one thing to write down a model, but another to fit it. For "power users" flexible software, like Nimble (de Valpine *et al.* 2017), which takes advantage of the flexibility and simplicity of the BUGS language, can be used to develop and extend these models. But for most ecologists it would be better to have bespoke software, which would mean developing packages like HMSC (Ovaskainen *et al.* 2017b) and gllvm (Niku *et al.* 2019) to fit models in the full framework.

We would not want to suggest that every ordination should use this framework: there will be times when the question being asked requires a different model. However, we feel that our framework is sufficiently flexible for many problems, and will build on both the modelling strengths developed from JSDMs and the visualisation and interpretation provided by the ordination world[1]. If extensions are needed, we would suggest that our model provides a starting point, both in terms of writing down the extended model and also in justifying why such an extension is needed. For example, when Björk *et al.* (2018a) used two ordinations to model two levels of sampling, a natural question is whether it would have been better to use one ordination, with sample and host species levels in the site effects, and with the same species loading. The answer to this question is not just one of modelling, it is also ecologically informative, i.e. about the extent to which species were responding to hosts, or to site effects.

[1] a world that is, of course, only two dimensional

Though the modelling framework suggested here is statistical in nature, this approach can serve to provide new insight into ecological theories on community assemblages. Further extensions of the framework towards spatiotemporal processes can serve to develop a theoretical model for how species interact in space and time, and how species-specific properties relate to site-specific properties, as in the case of functional traits.

The modelling of multispecies communities is being changed by the application of model sophisticated statistical techniques. Here we are suggesting that they can unify different fields, as they are modelling similar processes, and so can open up the fields to new questions. This is done both by moving ideas from one field to another (e.g. using ordination plots in JSDMs), and also by opening up new ways of analysing the data, with flexible models that can handle the problems associated with having many sites and species. To quote one of the world's great philosophers, let's go exploring (Watterso 1995).

Author Contributions

Usually a manuscript with a student and their supervisor comes about when the supervisor has a brilliant idea, but doesn't have time, so the student has to do all of the hard work, trying to find out if the idea actually works and doing the writing. In this manuscript, the roles were reversed. Neither author is totally sure how it happened.

References

Beissinger, S.R., Iknayan, K.J., Guillera-Arroita, G., Zipkin, E.F., Dorazio, R.M., Royle, J.A. & Kéry, M. (2016). Incorporating imperfect detection into joint models of communities: A response to warton et al. *Trends in Ecology & Evolution*, **31**, 736–737.

Björk, J.R., Hui, F.K.C., O'Hara, R.B. & Montoya, J.M. (2018a). Uncovering the drivers of host-associated microbiota with joint species distribution modelling. *Molecular Ecology*, **27**, 2714–2724.

Björk, J.R., Hui, F.K.C., O'Hara, R.B. & Montoya, J.M. (2018a). Uncovering the drivers of host-associated microbiota with joint species distribution modelling. *Molecular Ecology*, **27**, 2714–2724.

Björk, J.R., O'Hara, R.B., Ribes, M., Coma, R. & Montoya, J.M. (2018b). The dynamic core microbiome: Structure, dynamics and stability. *bioRxiv : the preprint server for biology*.

Blanchette, M.L. & Pearson, R.G. (2013). Dynamics of habitats and macroinvertebrate assemblages in rivers of the Australian dry tropics. *Freshwater Biology*, **58**, 742–757.

de Valpine, P., Turek, D., Paciorek, C., Anderson-Bergman, C., Temple Lang, D. & Bodik, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, **26**, 403–413.

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

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

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

Kissling, W.D., Dormann, C.F., Groeneveld, J., Hickler, T., Kühn, I., McInerny, G.J., Montoya, J.M., Römermann, C., Schiffers, K., Schurr, F.M., Singer, A., Svenning, J.-C., Zimmermann, N.E. & O'Hara, R.B. (2012). Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *Journal of Biogeography*, **39**, 2163–2178.

Mutshinda, C.M., O'Hara, R.B. & Woiwod, I.P. (2011). A multispecies perspective on ecological impacts of climatic forcing. *Journal of Animal Ecology*, **80**, 101–107.

Niku, J., Hui, F.K.C., Taskinen, S. & Warton, D.I. (2021). Analyzing environmental-trait interactions in ecological communities with fourth-corner latent variable models. *Environmetrics (London, Ont.)*, **32**, e2683.

Niku, J., Hui, F.K.C., Taskinen, S. & Warton, D.I. (2019). Gllvm - Fast analysis of multivariate abundance data with generalized linear latent variable models in R. *Methods in Ecology and Evolution*, **10**, 2173–2182.

O'Hara, R.B. & Sillanpää, M.J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis*, **4**, 85–117.

Ovaskainen, O., Tikhonov, G., Dunson, D., Grøtan, V., Engen, S., Sæther, B.-E. & Abrego, N. (2017a). How are species interactions structured in species-rich communities? A new method for analysing time-series data. *Proceedings of the Royal Society B: Biological Sciences*, **284**, 20170768.

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

Perrin, S.W., van der Veen, B., Golding, N. & Finstad, A.G. (2021). Modelling temperature-driven changes in species associations across freshwater communities. *Global Change Biology*, **n/a**.

Plummer, M. (2021). *Rjags: Bayesian graphical models using MCMC*.

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

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

ter Braak, C.J.F. (1985). Correspondence analysis of incidence and abundance data: Properties in terms of a unimodal response model. *Biometrics. Journal of the International Biometric Society*, **41**, 859–873.

ter Braak, C.J.F., Šmilauer, P. & Dray, S. (2018). Algorithms and biplots for double constrained correspondence analysis. *Environmental and Ecological Statistics*, **25**, 171–197.

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

Tikhonov, G., Abrego, N., Dunson, D. & Ovaskainen, O. (2017). Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context. *Methods in Ecology and Evolution*, **8**, 443–452.

Tobler, M.W., Kéry, M., Hui, F.K.C., Guillera-Arroita, G., Knaus, P. & Sattler, T. (2019). Joint species distribution models with species correlations and imperfect detection. *Ecology*, **100**, e02754.

Tredennick, A.T., Hooker, G., Ellner, S.P. & Adler, P.B. (2021). A practical guide to selecting models for exploration, inference, and prediction in ecology. *Ecology*, **102**, e03336.

van der Veen, B., Hui, F.K.C., Hovstad, K.A. & O'Hara, R.B. (2021a). Model-based ordination with constrained latent variables. *bioRxiv : the preprint server for biology*.

van der Veen, B., Hui, F.K.C., Hovstad, K.A., Solbu, E.B. & O'Hara, R.B. (2021b). Model-based ordination for species with unequal niche widths. *Methods in Ecology and Evolution*, **12**, 1288–1300.

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

Watterso, B. (1995). Final calvin and hobbes - last comic - by bill watterson for december 31, 1995 calvin and hobbes comic strip for december 31, 1995. URL https://www.gocomics.com/calvinandhobbes/1995/12/31 [accessed 12 November 2021]

Wilkinson, D.P., Golding, N., Guillera-Arroita, G., Tingley, R. & McCarthy, M.A. (2021). Defining and evaluating predictions of joint species distribution models. *Methods in Ecology and Evolution*, **12**, 394–404.

Wisz, M.S., Pottier, J., Kissling, W.D., Pellissier, L., Lenoir, J., Damgaard, C.F., Dormann, C.F., Forchhammer, M.C., Grytnes, J.-A., Guisan, A., Heikkinen, R.K., Høye, T.T., Kühn, I., Luoto, M., Maiorano, L., Nilsson, M.-C., Normand, S., Öckinger, E., Schmidt, N.M., Termansen, M., Timmermann, A., Wardle, D.A., Aastrup, P. & Svenning, J.-C. (2013). The role of biotic interactions in shaping distributions and realised assemblages of species: Implications for species distribution modelling. *Biological Reviews*, **88**, 15–30.