PeerJ

Generalized linear mixed models can detect unimodal species-environment relationships

Tahira Jamil[1,2] and Cajo J.F. ter Braak[1]

[1] Biometris, Wageningen University and Research Centre, Wageningen, The Netherlands
[2] Department of Mathematics, COMSATS Institute of Information Technology, Islamabad, Pakistan

ABSTRACT

Niche theory predicts that species occurrence and abundance show non-linear, unimodal relationships with respect to environmental gradients. Unimodal models, such as the Gaussian (logistic) model, are however more difficult to fit to data than linear ones, particularly in a multi-species context in ordination, with trait modulated response and when species phylogeny and species traits must be taken into account. Adding squared terms to a linear model is a possibility but gives uninterpretable parameters.

This paper explains why and when generalized linear mixed models, even without squared terms, can effectively analyse unimodal data and also presents a graphical tool and statistical test to test for unimodal response while fitting just the generalized linear mixed model. The R-code for this is supplied in Supplemental Information 1.

INTRODUCTION

Niche theory predicts that species occurrence and abundance show non-linear, unimodal relationships with respect to environmental gradients (*Austin, 1987*; *Palmer & Dixon, 1990*; *Whittaker, 1967*). Many studies fail to test for unimodal response (*Austin, 2007*). Thus straight-line relationships are often fitted without justification (e.g., *Gibson et al., 2004*). *Pollock, Morris & Vesk (2012)* propose a generalized linear mixed model for investigating trait modulation of the environmental response of a number of species. In their data unimodal response was said to be precluded, presumably as they examined relatively short environmental gradients. But might their method or a small modification thereof have worked for unimodal response? *Ives & Helmus (2011)* recently proposed phylogenetic generalized linear mixed models. Can these models usefully analyze unimodal response?

A similar question arises in community ecological ordination, a class of multivariate methods to analyze the occurrence and/or abundance of a set of species in a set of sites and resulting in a configuration of the sites in a factorial plane, the directions of which can be interpreted as latent environmental variables (*Jongman, ter Braak & van Tongeren, 1995*; *ter Braak & Prentice, 1988*; *Walker & Jackson, 2011*). Principal component analysis and

redundancy analysis are linear ordination methods whereas (detrended) correspondence analysis and canonical correspondence analysis are claimed to be able to analyze unimodal response (*ter Braak, 1985*; *ter Braak, 1986*). Nevertheless, (canonical) correspondence analysis is an eigen vector method and therefore inherently linear. This is most apparent in the reconstitution formula of (canonical) correspondence (*Greenacre, 1984*; *ter Braak & Verdonschot, 1995*). How can it be understood that these methods are able to model unimodal data but are inherently linear?

Some insight in this question is given by *Ihm & Van Groenewoud (1984)* and further worked out by *ter Braak (1987)* and *de Rooij (2007)* who showed the relationship between the unimodal model and a generalized (bi)linear model, also known as Goodman's RC model. The relationship can be used both ways. *Ihm & Van Groenewoud (1984)* use the relationship to justify their model B which is a (bi)linear model, for ecological ordination and *de Rooij (2007)* uses it to transform the linear predictor of the RC model into a quadratic form, with the graphical purpose to transform a vector representation or biplot to a distance representation that is supposed to be easier to interpret for naïve users of multivariate methods. Additional insight is given by *Zhu, Hastie & Walther (2005)* who introduced a weighted sample model to show the equivalence of canonical correspondence analysis and linear discriminant analysis.

In this paper we propose a graphical tool and statistical test to test for unimodal response while fitting just a generalized linear mixed model (GLMM) without squared terms. GLMMs are model-based, inferential statistical tools. GLMMs are very useful for describing the community patterns and are becoming popular in ecological and evolutionary studies (*Bolker et al., 2009*; *Ives & Helmus, 2011*; *Zuur et al., 2009*). Even when unimodality is detected using the proposed tool and test, we claim that GLMMs can effectively analyze unimodal data when the niche width is not very different among species and illustrate this claim by comparing the GLMM approach with an explicit unimodal model approach on data that show unimodal response. We focus on explicit environmental variables. In this case of direct gradient analysis, we can of course add the squares of environmental variables to the model (*ter Braak & Looman, 1986*) and test the statistical significance of the addition, but the argument extends to the trait modulated response (*Jamil et al., 2012*) and to latent variable models (*Walker & Jackson, 2011*) for which models without squared terms are easier to handle with software that is more widely available.

MATERIAL & METHODS

Generalized linear mixed models and unimodal response

For ease of exposition we start with a logistic linear mixed model for presence–absence data as example GLMM. The same approach can be followed for count data and loglinear models, which would relate to the RC model (*de Rooij, 2007*). Consider the logistic linear mixed model that relates the probability of occurrence p_{ij} of species j in site i to a quantitative environmental variable x_i by the formula

$$\text{logit}(p_{ij}) = \alpha_j + \beta_j x_i + \gamma_i \quad (i = 1, \ldots, n; j = 1, \ldots, m) \tag{1}$$

with α_j an intercept, β_j a slope and γ_i a site effect, which we all take as random parameters with normal distributions with zero mean and variances σ_α^2, σ_β^2 and σ_γ^2. In this random intercept, random slope model (*Gelman & Hill, 2007*; *Zuur et al., 2009*) it is prudent to have an additional parameter ρ for the correlation between the intercepts $\{\alpha_j\}$ and slopes $\{\beta_j\}$; otherwise the model would change by just centering the environmental variable. Inclusion of the random site effects $\{\gamma_i\}$ are a means to avoid pseudoreplication (*Hurlbert, 1984*) as they introduce correlation among species. These correlations were not modelled by *Pollock, Morris & Vesk (2012)* which makes their statistical tests liberal. The site effects may account for the size of the site, the fertility of the site or any other unknown factors that influence the probability of occurrence of all species in the site. The site effect γ_i will thus be expected to be related to the expected number of species in a site, that is to $\sum_j p_{ij}$ and, in terms of the data, to the number of species that is observed in a site, for short the site total, defined as $S_i = \sum_j y_{ij}$. The site total and the site effect are expected to have a monotonic positive relationship.

We now turn to one of the simplest unimodal curves for presence–absence, the Gaussian logistic curve (*Oksanen et al., 2001*; *ter Braak & Looman, 1986*)

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

with a_j a coefficient related to maximum probability of occurrence, u_j the species optimum and t_j the tolerance of species j. This model thus has a logistic form but is nonlinear in this parameterization. By expanding the quadratic term,

$$a_j - \frac{(x_i - u_j)^2}{2t_j^2} = a_j - \frac{1}{2t_j^2}x_i^2 - \frac{1}{2t_j^2}u_j^2 + \frac{1}{t_j^2}x_i u_j = \left(a_j - \frac{1}{2t_j^2}u_j^2\right) + \left(\frac{u_j}{t_j^2}\right)x_i - \frac{1}{2t_j^2}x_i^2 \tag{3}$$

it can be fitted to the data of each individual species by a generalized linear model (GLM) by using x and x^2 as predictors (*Jongman, ter Braak & van Tongeren, 1995*; *ter Braak & Looman, 1986*). The fit is unimodal if the regression coefficient of x^2 is negative.

Might we be able to model unimodal response even without the squared term? On assuming $t_j = t$ and setting

$$\alpha_j = a_j - \frac{1}{2t^2}u_j^2, \qquad \beta_j = \frac{u_j}{t^2} \quad \text{and} \quad \gamma_i = \frac{-1}{2t^2}x_i^2, \tag{4}$$

we obtain Eq. (1) again. If t would vary among species then Eq. (1) does not exactly hold because x_i^2/t_j^2 then also depends on j. With equal tolerances, unimodal response can thus be represented by a simple linear model with site effects and, as we propose, be fitted by a GLMM based on Eq. (1). The GLMM has additional normality assumptions. In case of unimodal response, the assumption that the site effects $\{\gamma_i\}$ are independent normal is false, as the site effects then depend on x_i through x_i^2/t^2. This will be the basis of our test on unimodal response in the next section. The site effects in Eq. (4) also have a nonzero mean, but that is not a problem, as the mean can be taken out and transferred to the intercepts (α_j).

The unimodal model with two or more environmental variables (*ter Braak & Prentice, 1988*) can similarly be rewritten as a simple linear model without squared terms if the tolerances are equal (Appendix 1). In conclusion, up to distributional assumptions, the GLMM (e.g., Eq. (1)) can be interpreted as a Gaussian logistic model with equal tolerances for the species.

A graphical tool and statistical test for unimodal response

Equation (4) suggests a graphical tool for detecting unimodal response and also a statistical test. The idea is to fit a GLMM to the binary data $\{y_{ij}\}$ with respect to the environmental variable with values $\{x_i\}$ ($i = 1, \ldots, n$). In the R package lme4 (*Bates, Maechler & Bolker, 2011*), the model can be fitted by

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

where y represents the vectorized response data while `sp` and `site` are factors indicating species and sites. The site effects $\{\gamma_i\}$ obtained from the fit are then plotted against the environmental variable $\{x_i\}$. There is an indication of unimodal response in terms of the species response with respect to the environmental variable x if this graph shows a n-shaped (as opposed to u-shaped) quadratic relationship. If the shape is not quadratic but curved, a transformation of x may improve it.

In the statistical test for unimodal response, the null model is the GLMM of Eq. (1) with, specifically, independent and normally distributed site effects. The alternative model is that the site effects depend quadratically on the environmental variable. As the site effects typically also depend on the site total S, a sensitive test on unimodal response is obtained by regressing the site effects on x, x^2 and S, according to the model formula

$$\gamma \sim x + x^2 + S. \tag{5}$$

There is evidence of unimodal response if the squared term is significant as judged by a z-test or, equivalently, an ANOVA test on its regression coefficient, the null model being $\gamma \sim x + S$. The R-code for making the graph and performing the test on unimodality is supplied in Supplemental Information 1.

Simulated data

In the first example series the procedure to simulate data is the following:

(1) Generate $n = 50$ values of an environmental variable x as a random sample from the uniform distribution, $x \sim U(-2, 2)$.

(2) Generate a vector u of length m from a uniform distribution $U(-\tau, \tau)$, where $\tau = 2 + t$, for a fixed value of t, to ensure that optima are also placed outside the sample range of x.

(3) Generate a vector a of length m drawn at random from the standard normal distribution.
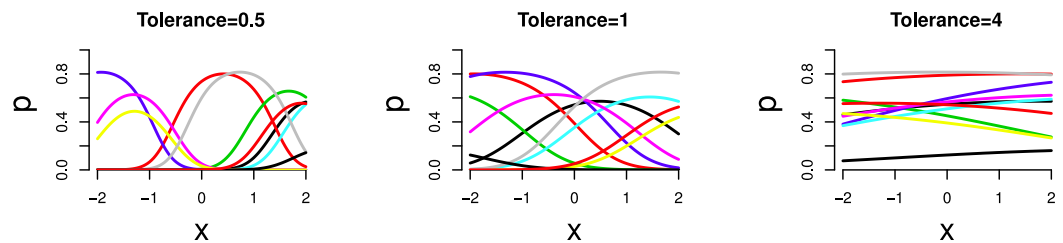
Figure 1 Simulated unimodal response (occurrence probability p_{ij} of a species at a site) against environmental variable x for a selection of species. For left to right: three tolerances ($t = 0.5$, 1 and 4).

(4) Generate binomial probabilities p_{ij} from the unimodal response curve

$$p_{ij} = \text{logit}^{-1}\left(a_j - \frac{(x_i - u_j)^2}{2t_j^2} \right) \tag{6}$$

and generate presence–absence data y_{ij} at random from a Bernoulli distribution with probability p_{ij} and $t_j = t$. We simulate data with constant tolerance in each data set for $m = 100$ species and vary t between data sets ($t = 0.5$, 1 and 4). Figure 1 indicates how the simulated species response curves look in each data set. We repeated the example with x and u having normal distributions instead ($x \sim N(0,1)$ and $u \sim N(0, t^2)$), but do not show the results as they were very similar to the uniformly distributed case.

In the second example series, we vary the tolerance lognormally among species, $t \sim \text{Log}\,N(0, \sigma)$ with $\sigma = 0.25$, 0.5 and 1 so that the median tolerance is 1. For $\sigma = 1$, the coefficient of variation is larger than 100%. In the third example series we vary the number of species ($m = 10$, 50 and 100) with $t = 1$. For the rest, the simulation is set up as in the first series. We also simulated data according to Eq. (1) with the strict assumptions of the GLMM of independent normal site effects.

Each dataset was characterized by beta diversity and length of gradient. The index of beta diversity is $\beta_w = T/S - 1$, where T is the total number of species, and S is the average number of species per site (*Whittaker, 1960*). Length of gradient, obtained by analyzing the data with detrended correspondence analysis (DCA), is expressed in standard deviation (SD) units (*Hill & Gauch, 1980*). Values greater than 4 are commonly taken to indicate unimodal response. Beta diversity was calculated using the asbio package (*Aho, 2011*) and DCA was performed in the *vegan* package (*Oksanen et al., 2011*), both in R.

Real data

We illustrate our method with three real data sets. The first is the Dune Meadow data (*Jongman, ter Braak & van Tongeren, 1995*). This is a small data set of 28 higher plants in 20 sites in a dune area in the Netherlands. Environmental variables, related to soil and management, were measured at each site; we use the variable Moisture.

The second data set includes the vegetation of the rising seashore on the island Skabbholmen in the Stockholm archipelago, eastern central Sweden (*Cramer & Hytteborn, 1987*) and is part of the Canoco package (*ter Braak & Smilauer, 1998*). The data set consists

of 63 sites sampled in both 1978 and 1984 and contains 68 species. The environmental variable is Elevation.

The third data set involves phytoplankton communities of 203 lakes located within four climate zones and associated measurements on various environmental variables and morphological species traits of 60 species (*Kruk et al., 2011*). We consider the environmental variable Temperature.

For each data set we fit a GLMM according to Eq. (1) with x the noted environmental variable, plot the resulting site effects against x and test for unimodal response at the 5% significance level explained by and below Eq. (5). For the seashore data, we analyzed 1978 and 1984 separately. We also compare the regression coefficient β_j as estimated by GLMM with the optimum u_j as obtained by explicitly fitting Eq. (2) using GLM (*ter Braak & Looman, 1986*) for species with a well-defined optimum, that is, the squared term of which has a z-ratio smaller than -1 in the GLM model. In the small Dune Meadow data set we used z-ratio < 0.

RESULTS

Simulated data

Figure 1 shows the simulated response curves in example series 1; the sampled range of the environmental variable is the range of x shown. With increasing tolerance the part of the curves that is sampled shows less unimodal response. This is expressed quantitatively in the length of gradient SD units which varies between about 1 SD (not so unimodal) to 6 SD (very unimodal); the beta diversity varies correspondingly between 1 and 5.

In Fig. 2, the site effects estimated by the GLMM analysis of each of the simulated data sets are plotted against the environmental variable. In all three series, site effects shows a clear quadratic relationship with the environmental variable except for large tolerance ($t = 4$) in the first series. Note the decreasing range of site effects as the tolerance increases in the first row of Fig. 2; for large tolerance, the site effects are close to zero. Nevertheless, the squared term in Eq. (5) was significant in all examples ($P < 0.001$) so that the method detects unimodal response even if it is moderate ($t = 4$). When the data are simulated according to Eq. (1) with the strict normality assumptions of the GLMM, the squared term was not judged significant more often than expected on the basis of Type I error of the test.

Figure 3 shows the relationship between the random slopes (β_j) estimated by GLMM and the true optima (u_j) in the three simulation series. The relation is positive as predicted by Eq. (4). The relationship is weaker the larger the tolerance (Fig. 3, first row). With tolerance varying across species, the relationship continues to hold true surprisingly well (Fig. 3, second row), except perhaps when the coefficient of variation of the tolerance is large ($>100\%$). The larger the number of species the clearer the predicted relationships (last row of Figs. 2 and 3).

Real data example

The site effects estimated by GLMM show a quadratic relationship with the noted environmental variable in each of the three data sets (Fig. 4 top row), with the least
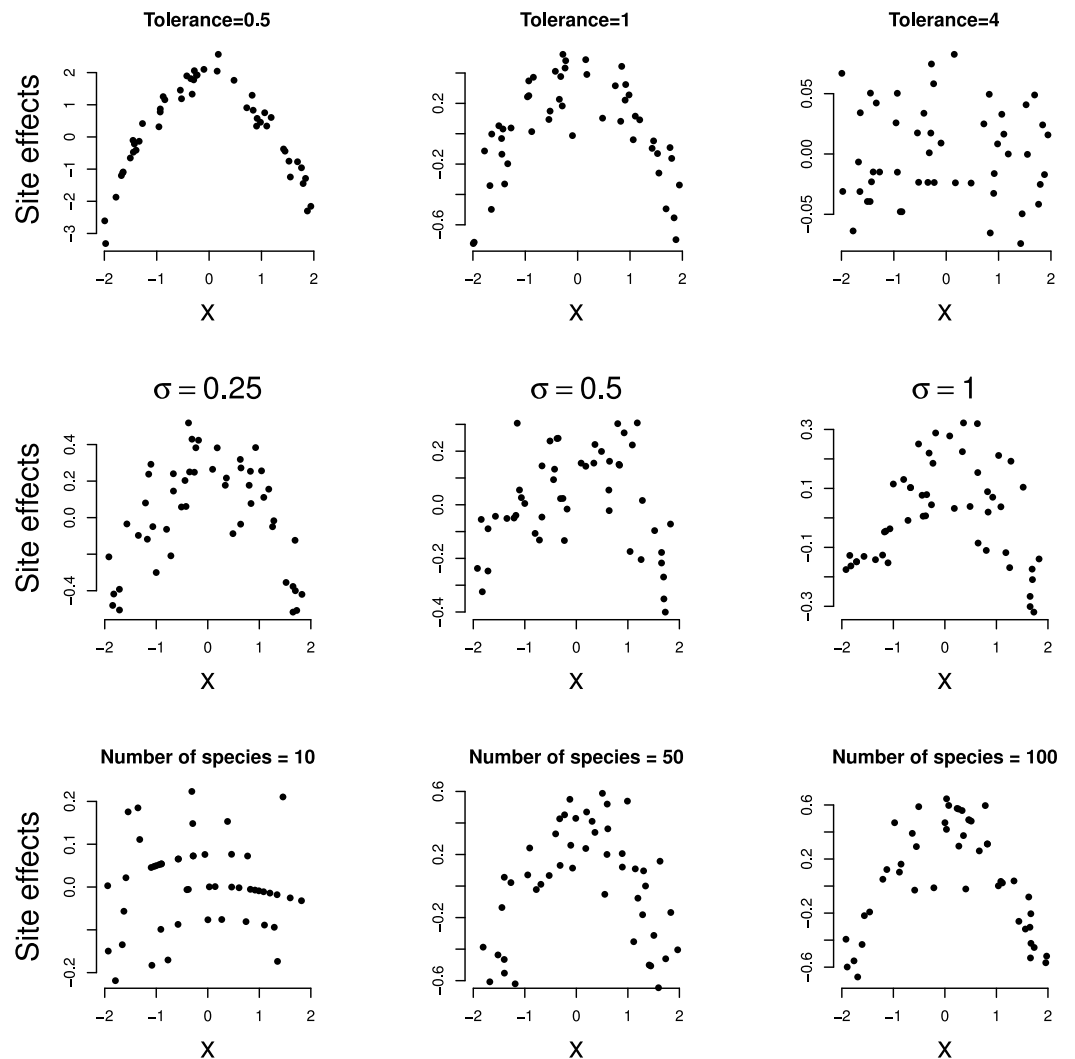
Figure 2 Simulated data: site effects as estimated by GLMM based on Eq. (1) plotted against environmental variable *x*. A quadratic relationship indicates a unimodal response. Rows: example series 1–3, columns: parameter varied.

unimodal response in the Dune Meadow data. Unimodal response is significant in all cases, as judged by our proposed significance test ($P < 0.001$) and the relationship between the random slopes (β_j) estimated by GLMM and the optima (u_j) obtained from a fit of the unimodal model of Eq. (1) is close to linear (Fig. 4, bottom row). In the Dune Meadow data there is one outlier for a species with a z-ratio close to 0. In the phytoplankton data, the species with similar low values for the optimum received differential values for the slope, but otherwise there is a good agreement.

DISCUSSION

Our analysis of multi-species data sets showed that a GLMM without squared terms but with site effects is able to detect unimodal response. The theory required equal tolerances among species, but the simulations showed remarkable robustness to this
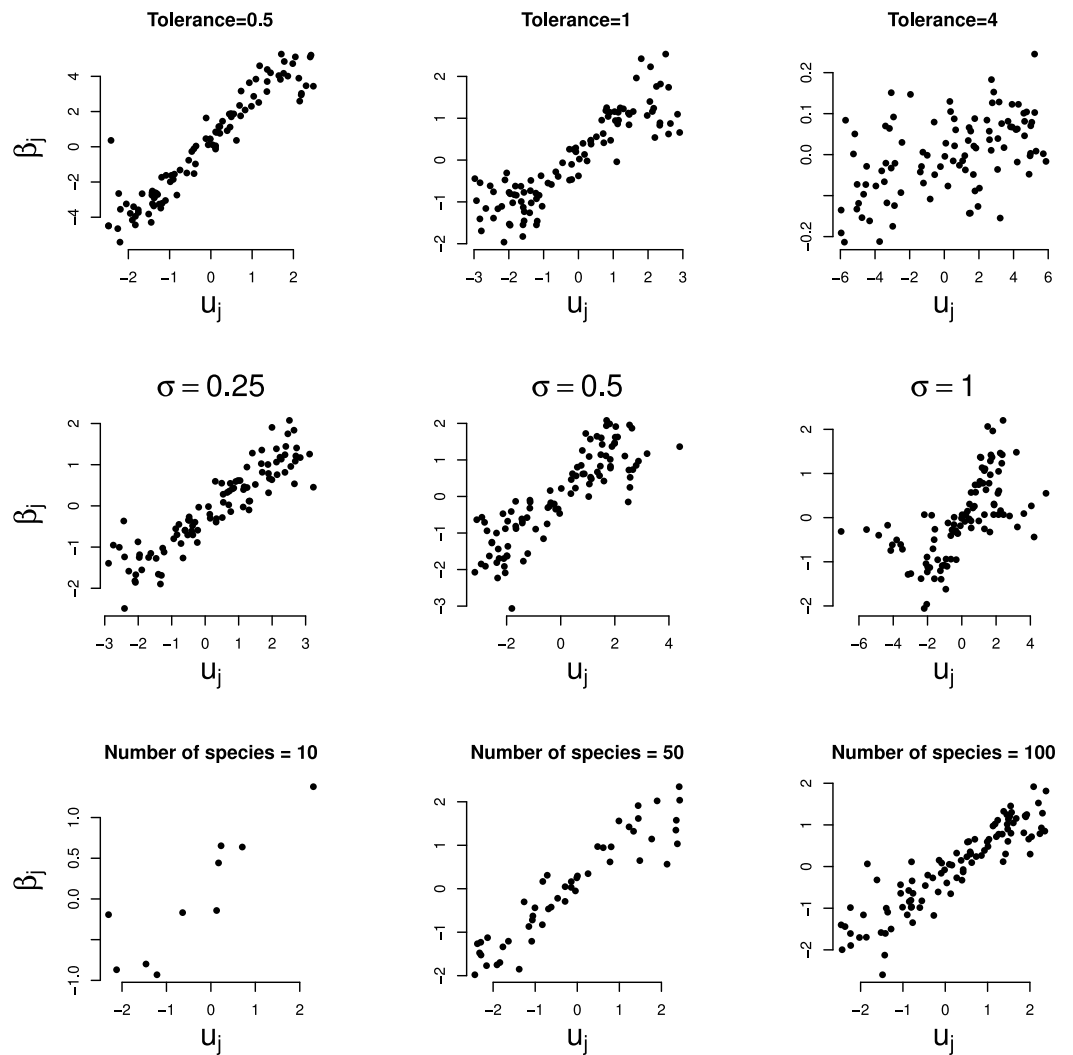
Figure 3 Simulated data: GLMM random slopes of Eq. (1) plotted against the true optima of the Gaussian logistic curve of Eq. (2). Rows: example series 1–3, columns: parameter varied. In the two bottom rows the (median) tolerance is 1.

assumption. With significant unimodal response, as judged by our test, the assumption of independence and normality of the site effects underlying the GLMM is clearly violated. This result can be interpreted in two ways. The first way is to try and adapt the model so that the assumptions are no longer (grossly) violated, for example, by extending the GLMM with an explicit square of the environmental variable as fixed effect, yielding

$$\text{logit}(p_{ij}) = \alpha_j + \beta_{1j} x_i + \beta_2 x_i^2 + \gamma_i \tag{7}$$

and then testing the normality assumption on the new site effects, for example by making a Q–Q plot (*Gelman & Hill, 2007*). The site effects γ_i obtained from the fit can then be plotted against the environmental variable to assess whether or not the unimodality is adequately modeled using quadratic terms. The extended model still assumes constant
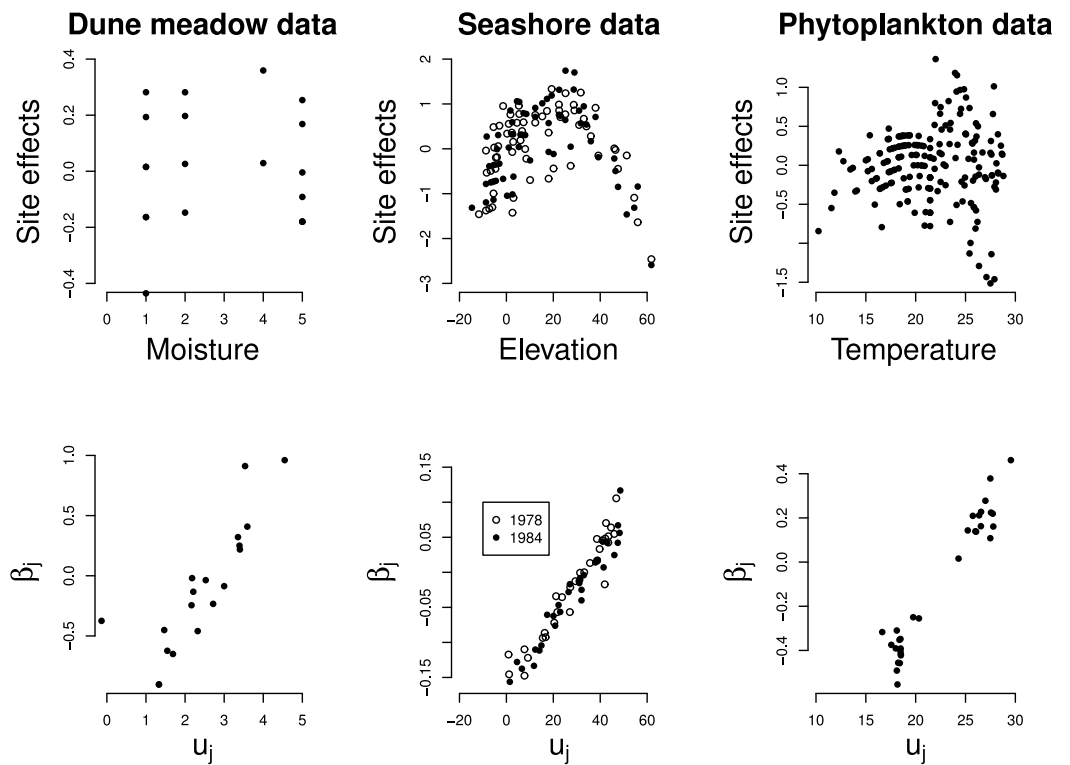
Figure 4 Real data: site effects (top) and the GLMM random slopes (bottom) plotted against the environmental variable *x* and the optima of the Gaussian logistic curve as estimated by GLM, respectively.

tolerance (as does Eq. (1)) as it can be rewritten as

$$\text{logit}(p_{ij}) = a_j - \frac{(x_i - u_j)^2}{2t^2} + \widetilde{\gamma}_i \tag{8}$$

with

$$t = 1/\sqrt{-2\beta_2}, \qquad u_j = t^2\beta_j, \qquad a_j = \alpha_j + \frac{1}{2t^2}u_j^2, \quad \text{and} \quad \widetilde{\gamma}_i = \gamma_i + \frac{1}{2t^2}x_i^2. \tag{9}$$

We can go one step further and test the assumption of equi-tolerance by adding the squared term x^2 also as a random (species-dependent) component to Eq. (7) and testing the significance of this extra variance component. In the R package `lme4` (*Bates, Maechler & Bolker, 2011*), the two models to compare are (with $xx = x^2$)

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

and

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

We are currently investigating robust ways of testing the equi-tolerance assumption on the basis of these two models, as the default ANOVA test has inflated type I error for small t.

The second way of interpretation is to conclude that the GLMM based on Eq. (1) is remarkably robust to the normality assumption on the site effects and the equi-tolerance assumption and that it can be used as a basis of more complicated models, such as the trait modulated response model (*Jamil et al., 2012*; *Pollock, Morris & Vesk, 2012*) and latent variable models (*Walker & Jackson, 2011*), even for unimodal response. In this way, the site effects are treated as nuisance parameters and the independence and normal distributions, needed for efficient computation, as a prior distribution that a posteriori may turn out to be false. *Cormont et al. (2011)* used this type of rationale to claim that their linear trait-environment method is well suited to analyze unimodal response.

CONCLUSIONS

Site effects in multi-species GLMM serve three purposes in ecological data, first to avoid pseudoreplication (*Hurlbert, 1984*), second to account for differences in species richness among sites (*Ives & Helmus, 2011*) and third (this paper) to allow for common nonlinear, unimodal response. The results of this paper imply that the phylogenetic generalized linear mixed models of *Ives & Helmus (2011)* are not in conflict with niche theory as they include random site effects and can thus deal with unimodal response.

Walker & Jackson (2011) used a latent variable approach to test for unimodal response. We tried their approach with the phytoplankton data, but failed to get an answer because the program for fitting the quadratic model crashed. We focused this paper on the easier task of detection unimodal response to measured environmental variables by using a GLMM without squared terms. To scale up to a latent variable approach we need factor analytic structure within a GLMM. This already exists for linear mixed models (*Thompson et al., 2003*; *Verbyla et al., 2003*) and it is a matter of time that it becomes standardly available for GLMM. This paper shows the utility of such factor analytic models in ecology (*Walker & Jackson, 2011*) if they allow additional random site effects.

A GLMM with terms that are linear in quantitative predictors is, of course, linear. But with random site effects included, GLMMs can detect and fit unimodal response, with the provision that the differences in niche widths among species is not too large. The application scope of GLMM in ecology is thus much broader than one might think at first glance.

ACKNOWLEDGEMENTS

Appendix 1

This appendix shows that similarity between the Gaussian logit models and GLMMs extends beyond the Gaussian logistic curve. The Gaussian logistic model can be extended to two environmental variables as follows (*ter Braak & Prentice, 1988*)

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

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

$$\alpha_j = a_j - \frac{1}{2}(d_1 u_{1j}^2 + d_2 u_{2j}^2 - 2d_{12j}u_{1j}u_{2j}),$$
$$\beta_{1j} = d_1 u_{1j} - d_{12j}u_{2j},$$
$$\beta_{2j} = d_2 u_{2j} - d_{12j}u_{1j},$$
$$\beta_{3j} = d_{12j},$$

and

$$\gamma_i = -\frac{1}{2}(d_1 x_{i1}^2 + d_2 x_{i2}^2)$$

we can write

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

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

ADDITIONAL INFORMATION AND DECLARATIONS

Competing Interests
CJFtB is an Academic Editor for PeerJ.

Author Contributions
- Tahira Jamil conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper.
- Cajo J.F. ter Braak conceived and designed the experiments, contributed reagents/materials/analysis tools, wrote the paper.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.95.

REFERENCES

Aho K. 2011. asbio: a collection of statistical tools for biologists. R package version 0.3-36. *Available at* http://cran.r-project.org (accessed 29 November 2011).

Austin MP. 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological Modelling* **200**:1–19 DOI 10.1016/j.ecolmodel.2006.07.005.

Austin MP. 1987. Models for the analysis of species' response to environmental gradients. *Plant Ecology* **69**:35–45 DOI 10.1007/BF00038685.

Bates D, Maechler M, Bolker B. 2011. lme4: linear mixed-effects models using S4 classes. R package version 0.999375-39. *Available at* http://cran.r-project.org (accessed 29 November 2011).

Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White JSS. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution* **24**:127–135 DOI 10.1016/j.tree.2008.10.008.

Cormont A, Vos CC, van Turnhout CAM, Foppen RPB, ter Braak CJF. 2011. Using life-history traits to explain bird population responses to increasing weather variability. *Climate Research* **49**:59–71 DOI 10.3354/cr01007.

Cramer W, Hytteborn H. 1987. The separation of fluctuation and long-term change in vegetation dynamics of a rising seashore. *Plant Ecology* **69**:157–167 DOI 10.1007/BF00038697.

de Rooij M. 2007. The distance perspective of generalized biadditive models: scalings and transformations. *Journal of Computational and Graphical Statistics* **16**:210–227 DOI 10.1198/106186007X180101.

Gelman A, Hill J. 2007. *Data analysis using regression and multilevel/hierarchical models.* Cambridge: Cambridge University Press.

Gibson LA, Wilson BA, Cahill DM, Hill J. 2004. Spatial prediction of rufous bristlebird habitat in a coastal heathland: a GIS-based approach. *Journal of Applied Ecology* **41**:213–223 DOI 10.1111/j.0021-8901.2004.00896.x.

Greenacre MJ. 1984. *Theory and applications of correspondence analysis.* London: Academic Press.

Hill MO, Gauch HG. 1980. Detrended correspondence analysis: An improved ordination technique. *Plant Ecology* **42**:47–58 DOI 10.1007/BF00048870.

Hurlbert SH. 1984. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* **54**:187–211 DOI 10.2307/1942661.

Ihm P, Van Groenewoud H. 1984. Correspondence analysis and Gaussian ordination. *COMPSTAT Lectures* **3**:5–60.

Ives AR, Helmus MR. 2011. Generalized linear mixed models for phylogenetic analyses of community structure. *Ecological Monographs* **81**:511–525 DOI 10.1890/10-1264.1.

Jamil T, Ozinga WA, Kleyer M, ter Braak CJF. 2012. Selecting traits that explain species–environment relationships: a generalized linear mixed model approach. *Journal of Vegetation Science* DOI 10.1111/j.1654-1103.2012.12036.x.

Jongman RHG, ter Braak CJF, van Tongeren OFR. 1995. *Data analysis in community and landscape ecology*. Cambridge: Cambridge University Press.

Kruk C, Peeters E, Van Nes EH, Huszar VLM, Costa LS, Scheffer M. 2011. Phytoplankton community composition can be predicted best in terms of morphological groups. *Limnology and Oceanography* **56**:110–118 DOI 10.4319/lo.2011.56.1.0110.

Oksanen J, Blanchet FG, Kindt R, Legendre P, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H. 2011. vegan: community Ecology Package. R package version 1.17-12. *Available at* http://cran.r-project.org (accessed 29 November 2011).

Oksanen J, Läärä E, Tolonen K, Warner BG. 2001. Confidence intervals for the optimum in the Gaussian response function. *Ecology* **82**:1191–1197 DOI 10.1890/0012-9658(2001)082[1191:CIFTOI]2.0.CO;2.

Palmer MW, Dixon PM. 1990. Small-scale environmental heterogeneity and the analysis of species distributions along gradients. *Journal of Vegetation Science* **1**:57–65 DOI 10.2307/3236053.

Pollock LJ, Morris WK, Vesk PA. 2012. The role of functional traits in species distributions revealed through a hierarchical model. *Ecography* **35**:716–725 DOI 10.1111/j.1600-0587.2011.07085.x.

Rue H, Held L. 2005. *Gaussian Markov random fields: theory and applications*. London: Chapman and Hall–CRC Press.

ter Braak CJF. 1985. Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. *Biometrics* **41**:859–873 DOI 10.2307/2530959.

ter Braak CJF. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* **67**:1167–1179 DOI 10.2307/1938672.

ter Braak CJF. 1987. Unimodal models to relate species to environment. PhD thesis, Wageningen University. *Available at* http://edepot.wur.nl/201452.

ter Braak CJF, Looman CWN. 1986. Weighted averaging, logistic regression and the Gaussian response model. *Vegetatio* **65**:3–11 DOI 10.1007/BF00032121.

ter Braak CJF, Prentice IC. 1988. A theory of gradient analysis. *Advances in Ecological Research* **18**:271–317 DOI 10.1016/S0065-2504(08)60183-X.

ter Braak CJF, Smilauer P. 1998. *CANOCO Reference Manual and User's Guide to Canoco for Windows: Software for Canonical Community Ordination (version 4)*. Ithaca: Microcomputer Power. *Available at* http://www.canoco.com.

ter Braak CJF, Verdonschot PFM. 1995. Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquatic Sciences* **57**:255–289 DOI 10.1007/BF00877430.

Thompson R, Cullis B, Smith A, Gilmour A. 2003. A sparse implementation of the average information algorithm for factor analytic and reduced rank variance models. *Australian & New Zealand Journal of Statistics* **45**:445–459 DOI 10.1111/1467-842X.00297.

Verbyla AP, Eckermann PJ, Thompson R, Cullis BR. 2003. The analysis of quantitative trait loci in multi-environment trials using a multiplicative mixed model. *Australian Journal of Agricultural Research* **54**:1395–1408 DOI 10.1071/AR02239.

Walker SC, Jackson DA. 2011. Random-effects ordination: describing and predicting multivariate correlations and co-occurrences. *Ecological Monographs* **81**:635–663 DOI 10.1890/11-0886.1.

Whittaker RH. 1960. Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs* **30**:279–338 DOI 10.2307/1943563.

Whittaker RH. 1967. Gradient analysis of vegetation. *Biological Reviews of the Cambridge Philosophical Society* **49**:207–264.

Zhu M, Hastie TJ, Walther G. 2005. Constrained ordination analysis with flexible response functions. *Ecological Modelling* **187**:524–536 DOI 10.1016/j.ecolmodel.2005.01.049.

Zuur AF, Ieno EN, Walker N, Saveliev AA, Smith GM. 2009. *Mixed effects models and extensions in ecology with R*. New York: Springer Verlag.