Generalized linear latent variable models for the analysis of multivariate abundance data

Sara Taskinen

University of Jyväskylä, Finland

Based on the joint work with

Francis Hui (ANU), John Ormerod (USyd), David Warton (UNSW),
Jenni Niku, Harri Högmander (JyU),
Emmanuela Daza Secco (JyU, SYKE), Kristian Meissner (SYKE)

Outline

Example

Classical multivariate analysis tools

Joint models for abundance

Computation

Simulation study

Example revisited

Summary

Example: Finnish peatland study

- Finnish environment institute is looking for new tools for ecological monitoring of peatlands.

- Former studies have shown that counts of amoeba species can be used for determining peatland condition (Daza Secco et al., 2016).

- As part of this study, we focus on the following research questions:
  1. Do amoeba species communities differ in terms of land use (natural, forestry, restored)?
  2. Can we find indicator species for different peatlands (natural, forestry, restored)?
  3. Do environmental variables (temperature and water pH) affect the community structure?

Data

- Six study sites located in the boreal zone of Central and Western Finland
  - Riihineva and Aittosuo (natural)
  - Lahnanen and Ruuskanlampi (forestry)
  - Aittoneva 60 and Aittoneva 80 (restored)
- 45 moss samples were taken from each sampling site.
- Amoeba species were identified and counted. Altogether 50 species were detected.
- Environmental variables (temperature and water pH) were measured from each sampling site.

Data matrix

```
       Cenacu Cencas Ceneco Cenpla Cycarc Triarc Trimin
site1    4544    802      0    267      0   1604      0
site2    2351      0      0    157      0   1881      0
site3    6415    802      0      0   1604    802      0
site4    6449      0      0      0      0    450      0
site5     948   2085      0      0   2843   1327      0
site6   11760    802      0      0    535   1336      0
site7    5957    526      0    175      0   1051      0
site8    5886      0      0      0      0      0      0
site9    4364      0      0      0      0    485      0
site10   2921      0      0    398      0    797      0
   .
   .
   .
```

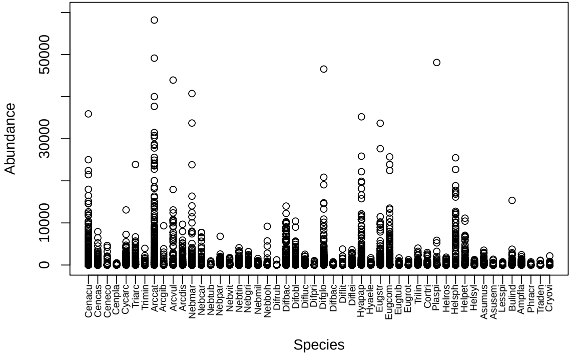

Figure: Abundances of $m = 50$ amoeba species recorded at $n = 270$ sites.

Classical multivariate analysis tools

- 'Algorithmic' multivariate analysis (Quinn and Keough, 2002) focuses on algorithms for ordination.
    - Aims at reducing data from many response variables to just two, so that sites can be plotted on a standard scatterplot to look for patterns between sites.
    - Methods are developed and implemented without directly accommodating the statistical properties (mean-variance relationship) of the data at hand.
- Species distribution modelling (see e.g. Elith and Leathwick, 2009).
    - Aims at predictive modeling and mapping the distribution of species and species diversity.
    - Less focus is put on correlations across species.

Joint models for abundance

- We analyze the data using a model-based approach. This allows us to specify a joint statistical model for abundance across many taxa.

- Model-based approaches allow us to
  - simultaneously explore interactions across taxa and the response of abundance to environmental variables,
  - explicitly account for key statistical properties of the data,
  - use residual analysis tools for model checking,
  - use model selection tools to choose the most appropriate model for data at hand,
  - use the standard tools developed for statistical inference.

Generalized linear models (GLM)

- The models we use are extensions of Generalized linear models (McCullagh and Nelder, 1989), widely used to model the impact of environmental predictors, $\boldsymbol{x}_i$, $i = 1, \ldots, n$, on abundance of one species, y_i, $i = 1, \ldots, n$.
- In GLM the mean response, denoted by $\mu_i = E(y_i)$ is assumed to be

  $$g(\mu_i) = \beta_0 + \boldsymbol{x}_i' \boldsymbol{\beta},$$

  where $g(\cdot)$ is a known link function, β_0 is an intercept and $\boldsymbol{\beta}$ is a vector of regression coefficients related to measured environmental covariates.

Generalized linear mixed models (GLMM)

- A joint model for abundance, y_{ij}, $i = 1, \dots, n$, $j = 1, \dots, m$, requires the inclusion of random effects, hence some form of mixed model, to capture correlation in abundance across taxa.

- A complicated way to incorporate correlation is to introduce it directly via a multivariate random effect applied to each sample, to form a multivariate generalized linear mixed model (Breslow and Clayton, 1993)

  $$g(\mu_{ij}) = \alpha_i + \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j + u_{ij},$$

  where α_i and β_{0j} denote row effects and species-specific intercepts, respectively, $\boldsymbol{\beta}_j$ are coefficient vectors related to the environmental covariates and $\boldsymbol{u}_i = (u_{i1}, \dots, u_{im})' \sim N(\mathbf{0}, \mathbf{\Sigma})$.

Generalized linear latent variable models (GLLVM)

- A flexible way to incorporate correlation is to regress the mean response μ_{ij} against a vector of $d \ll m$ unknown latent variables, $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{id})'$, along with covariates.

- This forms a multivariate generalized linear latent variable model (Moustaki and Knott, 2000), where

$$g(\mu_{ij}) = \alpha_i + \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j + \boldsymbol{u}_i'\gamma_j,$$

  where $\boldsymbol{u}_i \sim N(\mathbf{0}, \boldsymbol{I}_d)$ and $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jd})'$ are coefficients which quantify how each species response is related to the latent variable.

Generalized linear latent variable models (GLLVM)

- The term $\boldsymbol{u}_i'\gamma_j$ now captures the correlation across species, and the number of latent variables (d) controls model complexity.

- Latent variable can be used to produce an ordination plot. If $d = 2$, the latent variable value $\boldsymbol{u}_i$ is a pair of coordinates representing the position of the site i in a two-dimensional ordination (Hui et al., 2015).

- The coefficients γ_j can be added to the ordination giving an indication of how species composition differs across sites.

Computation

- Write $\boldsymbol{Y} = (\boldsymbol{y}_1 \cdots \boldsymbol{y}_n)'$ for a $n \times m$ response matrix and collect all model parameters into a vector $\boldsymbol{\Psi}$.

- We estimate the model parameters using the maximum likelihood method, that is, we find such $\boldsymbol{\Psi}$ which maximizes

$$L(\boldsymbol{\Psi}) = \prod_{i=1}^{n} f(\boldsymbol{y}_i; \boldsymbol{\Psi}).$$

- For GLLVMs, the marginal density function of $\boldsymbol{y}_i$ is given by

$$f(\boldsymbol{y}_i, \boldsymbol{\Psi}) = \int_{\mathbb{R}^d} \prod_{j=1}^{m} f(y_{ij}|\boldsymbol{u}_i; \boldsymbol{\Psi}) h(\boldsymbol{u}_i) d\boldsymbol{u}_i,$$

where $h(\cdot)$ is the density of d-variate standard normal distribution.

Computation

- As the marginal likelihood function involves a d-dimensional integral, which cannot be solved analytically, numerical approximation methods are needed.
- Methods available in the literature include
  - Gauss-Hermite (GH) quadrature (Moustaki, 1996; Moustaki and Knott, 2000) for mixtures of binary and normal responses,
  - Adaptive Gauss-Hermite (AGH) quadrature (Rabe-Hesketh et al., 2002) for normal, binomial, gamma and Poisson distributed responses,
  - Laplace approximation (Huber et al., 2004; Bianconcini and Cagnone, 2012) for responses from general exponential family,
  - MCEM (Sammel at al., 1997) for mixtures of binary and normal responses.

Computation

- Less methods are available for overdispersed count data.
- Recent contributions for R-software include
  - EM-algorithm (Hui et al., 2014)
  - MCMC (Hui, 2015)
  - Laplace approximation (Niku et al., 2016a)
  - Variational approximation (Hui et al., 2016b)

Variational approximation (VA) method

- Using VA method it is possible to construct a more tractable (potentially closed form) approximation to intractable likelihood.
- VA methods are popular for approximating posterior distributions in high dimensional Bayesian modelling.
- Ormerod and Wand (2012) used VA method to overcome the problems in integration in maximum likelihood estimation of generalized linear mixed models.

Variational approximation (VA) method

- Let $q(\cdot)$ be an arbitrary density function on $\mathbb{R}^d$. The log-likelihood can then be written as

$$
\begin{aligned}
l(\boldsymbol{\Psi}) = \log f(\boldsymbol{y};\boldsymbol{\Psi}) \int_{\mathbb{R}^d} q(\boldsymbol{u}) d\boldsymbol{u} &= \int_{\mathbb{R}^d} \log\left(\frac{f(\boldsymbol{y},\boldsymbol{u};\boldsymbol{\Psi})/q(\boldsymbol{u})}{f(\boldsymbol{u}|\boldsymbol{y};\boldsymbol{\Psi})/q(\boldsymbol{u})}\right) q(\boldsymbol{u}) d\boldsymbol{u} \\
&= \int_{\mathbb{R}^d} \log\left(\frac{f(\boldsymbol{y},\boldsymbol{u};\boldsymbol{\Psi})}{q(\boldsymbol{u})}\right) q(\boldsymbol{u}) d\boldsymbol{u} + \int_{\mathbb{R}^d} \log\left(\frac{q(\boldsymbol{u})}{f(\boldsymbol{u}|\boldsymbol{y};\boldsymbol{\Psi})}\right) q(\boldsymbol{u}) d\boldsymbol{u}.
\end{aligned}
$$

- The last term is the Kullback-Leibler distance between $q(\boldsymbol{u})$ and $f(\boldsymbol{u}|\boldsymbol{y})$. Since this is always nonnegative, we get

$$
l(\boldsymbol{\Psi}) \geq \int_{\mathbb{R}^d} \log\left(\frac{f(\boldsymbol{y},\boldsymbol{u};\boldsymbol{\Psi})}{q(\boldsymbol{u})}\right) q(\boldsymbol{u}) d\boldsymbol{u}. \tag{1}
$$

- Substitution of $q(\boldsymbol{u}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Lambda})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ are called the variational parameters, into (1) gives a closed form lower bound.

- Estimation of the GLLVM is performed by maximizing the VA log-likelihood simultaneously over the variational parameters and model parameters.

- For the analysis of model parameters, the approximate asymptotic standard errors may be obtained using the observed information matrix

$$\boldsymbol{I}(\hat{\boldsymbol{\Psi}}, \hat{\mu}, \hat{\boldsymbol{\Lambda}}) = -\left\{ \frac{\partial^2 \underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})}{\partial(\boldsymbol{\Psi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \partial(\boldsymbol{\Psi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})^T} \right\}_{\hat{\boldsymbol{\Psi}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Lambda}}}.$$

- The variational parameter estimates, $\hat{\boldsymbol{\mu}}_i$ provide appropriate approximations to best predictors of $\boldsymbol{u}_i$ (BP), and $\hat{\boldsymbol{\Lambda}}_i$ can be used to measure their variability (Ormerod and Wand (2010)).

VA log-likelihood for Poisson-Gamma model

- To handle overdispersed counts in the context of GLLVMs, we use a multiplicative Poisson-Gamma model with log link function, that is,

  $$f(y_{ij}|\nu_{ij}, \boldsymbol{u}_i, \boldsymbol{\Psi}) = \exp(-\nu_{ij})(\nu_{ij})^{y_{ij}}/y_{ij}!, \quad \nu_{ij} \sim \text{Gamma}(\phi_j, \phi_j/\mu_{ij})$$

  and $\log(\mu_{ij}) = \eta_{ij} = \alpha_i + \beta_{0j} + \boldsymbol{x}_i^T\boldsymbol{\beta}_j + \boldsymbol{u}_i^T\boldsymbol{\gamma}_j$.
- The parameterization produces the same quadratic mean-variance relationship as the negative binomial distribution, that is, $\text{Var}(y_{ij}) = \mu_{ij} + \mu_{ij}^2/\phi_j$, where ϕ_j is the dispersion parameter.
- Also a fully closed form VA log-likelihood is obtained.

VA log-likelihood for Poisson-Gamma model

Theorem

The VA log-likelihood for Poisson-Gamma GLLVM with log link function is given by the following expression

$$\begin{aligned}\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\Lambda}, \boldsymbol{\mu}) = \sum_{i=1}^{n} \sum_{j=1}^{m} \Bigg(& y_{ij}\left(\tilde{\eta}_{ij} - \frac{1}{2}\gamma_j^T \boldsymbol{\Lambda}_i \gamma_j\right) \\ & - (y_{ij} + \phi_j) \log\left\{\phi_j + \exp\left(\tilde{\eta}_{ij} - \frac{1}{2}\gamma_j^T \boldsymbol{\Lambda}_i \gamma_j\right)\right\} \\ & + \log\Gamma(y_{ij} + \phi_j) - \frac{\phi_j}{2}\gamma_j^T \boldsymbol{\Lambda}_i \gamma_j \Bigg) + n\{\phi_j \log(\phi_j) - \log\Gamma(\phi_j)\} \\ & + \frac{1}{2}\sum_{i=1}^{n}(\log\det(\boldsymbol{\Lambda}_i) - tr(\boldsymbol{\Lambda}_i) - \boldsymbol{\mu}_i^T \boldsymbol{\mu}_i),\end{aligned}$$

where $\tilde{\eta}_{ij} = \alpha_i + \beta_{0j} + \boldsymbol{x}_i^T \boldsymbol{\beta}_j + \boldsymbol{\mu}_i^T \boldsymbol{\gamma}_j$, *and all other quantities that are constant with respect to the parameters have been omitted.*

Simulation study

- $K = 200$ random samples were generated according to the Poisson-Gamma model using different sample sizes and dimensions.

- We cosidered simple GLLVM model without covariates and site effects, that is,

$$g(\mu_{ij}) = \beta_{0j} + \boldsymbol{u}_i'\gamma_j,$$

  where $\boldsymbol{\beta}_0 = (-1, \ldots, -1, 1, \ldots, 1)$, true latent variables $\boldsymbol{u}_i$ were generated from the mixture of bivariate normal distributions and the elements of γ_j were generated from the uniform distribution $U(-2, 2)$.

- We compared the results based on variational approximation method to those given by Laplace's method (Niku et al., 2016a).

Simulation study

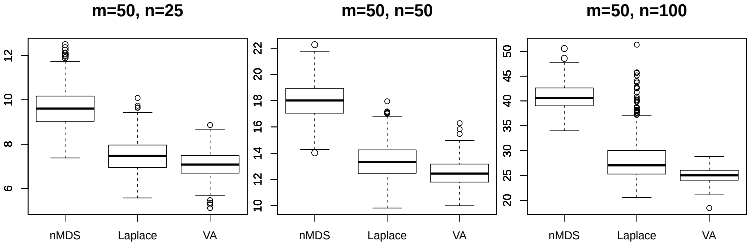

Figure: To evaluate the performance of predicted latent variables, $\boldsymbol{u}_i$, the procrustes errors between the predicted and true parameter values were computed (Bartholomew et al., 2011). Non-metric multidimensional scaling was added in comparisons as a classical ordination method.

Simulation study

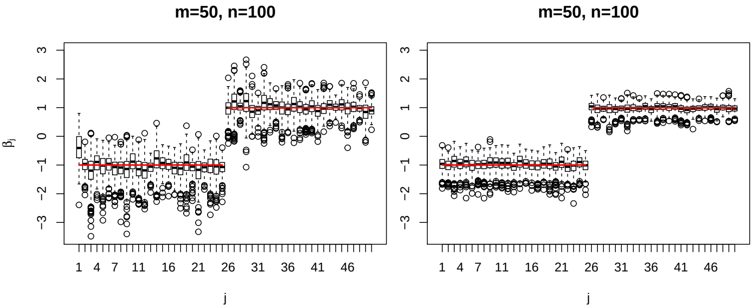

Figure: Boxplots of estimated regression coefficients, $\hat{\beta}_j$, and true values for β_j (red lines). The estimation was done using Laplace's method (left) and Variational approximation method (right).

Simulation study

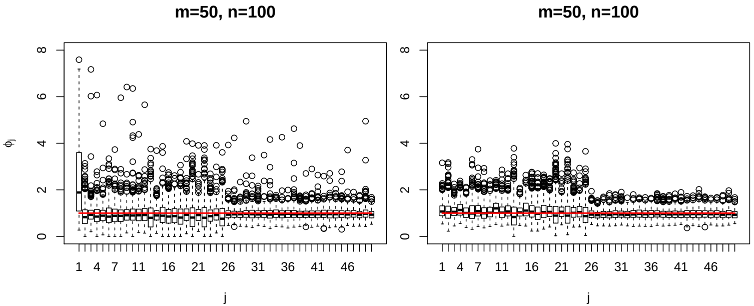

Figure: Boxplots of estimated dispersion parameters, $\hat{\phi}_j$, and true values for ϕ_j (red lines). The estimation was done using Laplace's method (left) and Variational approximation method (right).

Computation times

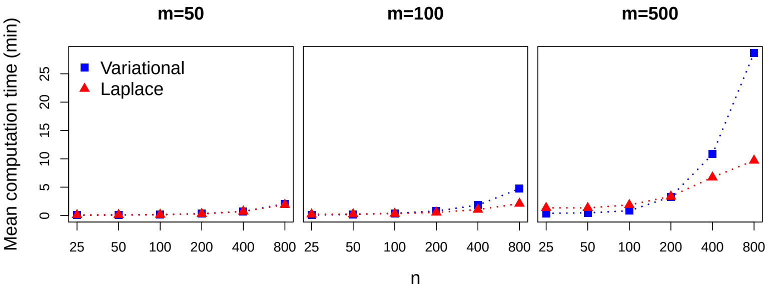

Figure: Mean computation times (in minutes) when GLLVMs with two latent variables were fitted using variational approximation method and Laplace approximation method.

Computation times

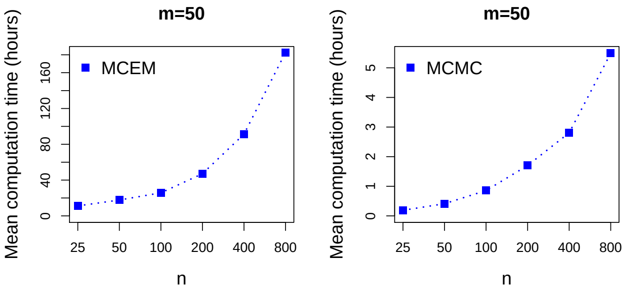

Figure: Mean computation times (in hours) when GLLVMs with two latent variables were fitted using EM algorithm utilizing Monte Carlo integration at E-step and MCMC method.

Example revisited: Finnish peatland study

- Consider the amoeba dataset with $n = 270$ sites and $m = 50$ species.
- To visualize the main trends between different sampling sites in terms of their species composition we fitted a latent variable model with two latent variables (model-based ordination method).
- We used model selection tools (BIC) and chose a model which assumes Poisson-Gamma distributions (over Poisson, ZIP and ZIP-NB) for responses.
- Model was fitted using VA method and predicted latent variables (BPs) were plotted on a standard scatterplot to look for patterns between sites.

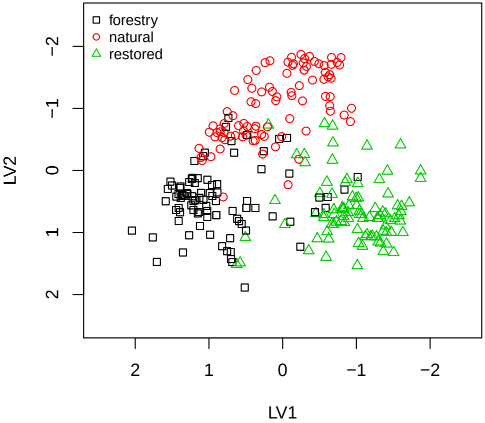

Figure: Predicted latent variables for amoeba dataset. Sites close to each other are similar in terms of their species composition.

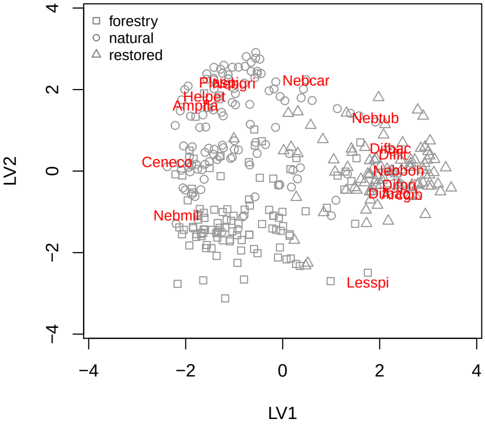

Figure: Model-based biplot for amoeba dataset. 15 species with the largest factor loadings (in terms of distance from the origin) are printed. Species in the same direction and far from the origin are highly correlated.

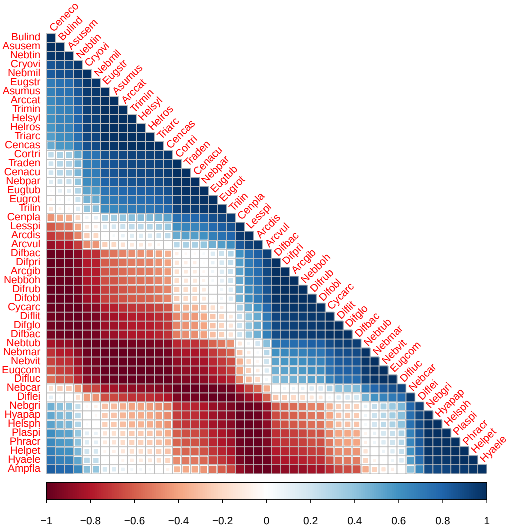

Figure: Color plot of the correlation matrix for amoeba dataset to show which species show high/low correlations.

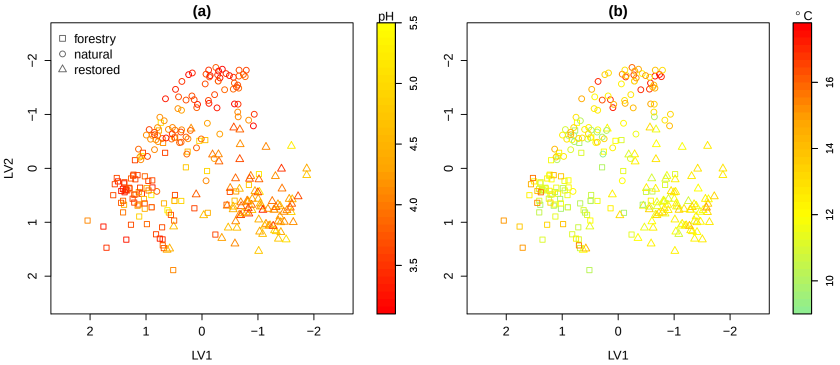

Figure: Predicted latent variables for amoeba dataset. Sites are shown in colors indexed by the value of (a) water pH and (b) temperature.

Summary

- GLLVMs allow us to specify a statistical model for abundances jointly across many taxa, to simultaneously explore interactions across taxa and the response of abundance to environmental variables.

- Advantages of model-based approaches include: residual analysis tools, model selection tools and methods for formal statistical inference.

- Fast estimation methods for GLLVMs are available for the most common types of responses in ecological studies: presence-absence records, overdispersed species counts, biomass (non-negative, continuous data often with large number of zeros), and percent cover data.

References

Elith, J. and Leathwick, J.R. (2009). Species distribution models: ecological explanation and prediction across space and time *Annual Review of Ecology, Evolution, and Systematics*, **40**, 677–697.

Hui, F.K.C. (2014). Boral: R package version 0.6.

Hui, F.K.C., Taskinen, S., Pledger, S., Foster, S.D. and 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. and Taskinen S. (2016b). Variational approximations for generalized linear latent variable models. *Journal of Computational and Graphical Statistics*, in print.

Moustaki, I. and Knott, M. (2000). Generalized latent trait models. *Psychometrika*, **65**, 391–411.

Niku, J., Warton, D.I., Hui, F.K.C. and Taskinen, S. (2016a). Generalized linear latent variable models for multivariate abundance data in ecology, manuscript.

Quinn, G.G.P. and Keough, M.J. (2002) *Experimental Design and Data Analysis for Biologists.* Cambridge University Press, Cambridge, UK.

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