Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data

JAMES S. CLARK[1,3], DIANA NEMERGUT[4], BIJAN SEYEDNASROLLAH[1], PHILLIP J. TURNER[2], and STACY ZHANG[2]

November 8, 2016

[1] *Nicholas School of the Environment, Duke University, Durham NC 27708, USA*

[2] *Division of Marine Science and Conservation, Nicholas School of the Environment, Duke University, Beaufort, NC 28516, USA*

[3] *Department of Statistical Science, Duke University, Durham, NC 27708-0251*

[4] *Department of Biology, Duke University, Durham, NC 27708*

Running head: Generalized joint attribute modeling

Abstract

Probabilistic forecasts of species distribution and abundance require models that accommodate the range of ecological data, including a joint distribution of multiple species based on combinations of continuous and discrete observations, mostly zeros. We develop a generalized joint attribute model (GJAM), a probabilistic framework that readily applies to data that are combinations of presence-absence, ordinal, continuous, discrete, composition, zero-inflated, and censored. It does so as a joint distributions over all species providing inference on sensitivity to input variables, correlations between species on the data scale, prediction, sensitivity analysis, definition of community structure, and missing data imputation. GJAM applications illustrate flexibility to the range of species-abundance data. Applications to forest inventory demonstrate species relationships responding as a community to environmental variables. It shows that the environment can be inverse predicted from the joint distribution of species. Application to microbiome data demonstrates how inverse prediction in the GJAM framework accelerates variable selection, by isolating effects of each input variable's influence across all species.

Keywords: community structure, categorical data, composition data, Generalized Joint Attribute Model, hierarchical model, Joint Species Distribution model, microbiome data, ordinal data, presence-absence, trait data

Introduction

Efforts to explain and predict biodiversity (e.g., Iversen and Prasad 1998; Ferrier et al. 2002; Guisan and Thuiller 2005; Gelfand et al. 2006; Araujo, and Luoto 2007; Botkin et al. 2007; Chakraborty et al. 2010; Benito et al. 2013, Booth et al. 2014) confront three challenges summarized in our title. First, *median-zero* refers to the fact that most of the values in species-abundance data sets are typically zero (Fig. 1b, c). Second, species are not independent and thus models must be *multivariate*. Finally, data may be continuous (density, basal area, biomass), discrete (presence/absence, counts), censored (detection limits, intervals, maximum values), composition (proportional of a total), nominal, and ordinal; such *multifarious* combinations of observations are not described by standard distributions. We describe generalized joint attribute modeling (GJAM) to address this challenge, providing a common framework for synthesis of ecological attribute and abundance data, both for estimating responses to the environmental and for prediction.

GJAM is motivated by the difficulties faced by all species distribution models (SDMs), including joint species distributions models (JSDMs)(Clark et al. 2014, Pollock et al. 2014) and predictive trait models (PTMs) (Clark 2016). SDMs and JSDMs omit much of the information contained in field data, where abundances and attributes are often documented in multifarious ways. Some species groups are counted. Those not easily measured are recorded in ordinal categories, such as ’rare’, ’moderate’, and ’abundant’. Presence-absence of a predator, pathogen, or mutualist might be recorded. Attributes such as body condition, infection status, and herbivore damage can be included. Even condition of a sample plot can be relevant. For example, grazer abundance might be observed together with evidence for plot-level grazing damage, as ordinal scores (’none’ to ’severe’) or categorical (nominal) categories. How would a model combine insect counts of multiple species from pitfall traps with herbaceous cover? Or fishing returns with presence-absence by-catch of threatened species? Or microbiome data with host condition and abundance (Fig. 1a, b)? All of these variables are ’responses’, not predictors–they are just as random as abundance values, both affecting and responding to other variables. All are recorded on different scales. We introduce the term generalized joint attribute model (GJAM) for the model that accommodates these attributes jointly.

The challenges of multifarious data may account for two tendencies in the SDM literature, i) to model on a transformed scale that is different from the data (e.g., a non-linear link function) and ii) to model something other than what was observed, most often substituting presence-absence for observations that come from many scales. Although several JSDMs apply to abundance data (Latimer et al. 2009; Thorson et al. 2015), and one applies to combined presence-absence and continuous abundance data (Clark et al. 2014), most assume presence-absence (Finley et al. 2009; Ovaskainen et al. 2011; Ovaskainen and Soininen 2011; Pollock et al. 2014; Harris 2015), even when data are not collected this way. The question becomes, do these modeling decisions affect inference and prediction?

First, the covariance matrix estimated in a hierarchical JSDM with non-linear link functions (Finley et al. 2009; Ovaskainen et al. 2011; Ovaskainen and Soininen 2011; Pollock et al. 2014; Thorson et al. 2015) is not estimated on the data scale and thus is not to be interpreted as a covariance between species abundances. When response variables are continuous and covary, their dependence structure is most efficiently modeled with a covariance matrix. However, many ecological data types are discrete (counts, ordinal scores, zeros, censored intervals). A covariance matrix can still be used in models of such data if it is introduced at a first stage of a hierarchical model, provided there is a non-linear link function to data. For example, a generalized linear model (GLM) can specify a Poisson distribution for counts, $y_{is} \sim Poi(\lambda_{is})$ of species s in observation i. This model for discrete counts does not admit a covariance matrix. The intensity λ_{is} is continuous, but unless there is a scale transformation, models for it too do not admit a covariance, because λ is constrained to positive values. The log transformation, or *link function*, introduces a new issue that is not widely appreciated, the fact that covariance cannot be interpreted on the scale of the observations y_{is}. Whereas intensity λ_{is} has the transparent interpretation on the same scale as the counts themselves, the covariance on the log scale does not (Fig. 2*a*). Then too, the explanatory variables subjected to non-linearity transformation also no longer have the transparent interpretations of 'main effect' and 'interaction'. On the transformed scale, all variables are part of interactions imposed by the form of the link function. If a sample contained multifarious data, complications would compound as each type of observation might require a different link function to allow for the second-stage continuous model that includes covariance. If it is already hard to attach meaning to covariance on the log scale, how can we interpret covariance structure where some responses are log scale and others logit scale (Fig. 2*b*)?

Non-linear link functions are generally not motivated by theory. A log link might be used because it accommodates an increase in variance with abundance. Mean-variance relationships are important to consider, but model adequacy is generally evaluated on the basis of residual errors or data prediction (e.g., Ver Hoef and Boveng 2007, Warton et al. 2012, Hui et al. 2015) rather than theory. Non-linear link functions can arise naturally when a likelihood function is written in exponential family form. However, models on the observation scale could also be valuable for many applications, particularly when observations on different scales are combined. They have transparent interpretation.

The second tendency, to substitute presence-absence models for data collected in other ways has not been evaluated for a joint distribution of species. When a study changes the observations, the loss of information (e.g., when abundance on many scales is reduced to 'presence') should affect estimates. The question is, how much?

If collapsing abundance to presence-absence or changing the data in other ways might come at a cost, why is it so often done? The consequences are not discussed in the literature and may be unrecognized. Without a GJAM, the effects demonstrated here would be hard to quantify, due to the different link functions used for presence-absence and abundance data. There has been little attention to the challenge posed by multifarious data.

The problem of zeros in species abundance data has been discussed in the context of univariate models (e.g., Martin et al. 2005). For count data, Poisson, negative binomial, and even hyper-zero-inflated models perform poorly when the fraction of zeros approaches 50% (Ghosh et al. 2012, Clark and Gelfand 2016). In many ecological data sets zeros can often exceed $> 90\%$ of all observations (Fig. 1), and the traditional solutions are limited. And again, presence-absence models cannot accommodate any species that are present in all samples. In joint models the challenge of overwhelming zeros must be confronted with models that also admit multifarious data.

The need for a model that allows flexibility for continuous, discrete, ordinal, and composition data, with censoring and zero inflation motivates a GJAM. We describe a synthetic framework for observations of many types, modeling the data on the scale they are collected, imposing a reference scale only for data that have none (e.g., presence-absence). The coefficients and species correlations in GJAM are interpretable on the observation scale.

An important extension of GJAM involves an expanded role for prediction. Objectives of SDM studies most often concern community-level variables, such as species richness, diversity, or biomass (Elith et al. 2006, Ferrier et al. 2002, Baselga and Araujo 2010, Guisan and Rahbek 2011, Mokany and Ferrier 2011, Mokany et al. 2011, 2012). Formal predictive modeling is not possible from SDMs fitted to species independently, requiring an informal approach that omits relationships between species (e.g. Calebrese et al. 2014). Beyond showing the value of in-sample and out-of-sample prediction to verify that GJAM applies to the many data types and species responses jointly, we go further. Inverse prediction provides a composite estimate of environmental importance for the entire community (Clark et al. 2011, 2013). It opens new options for predicting the environment from species, because it combines information from all species in a synthetic prediction with full uncertainty. Predictive distributions allow us to explore community structure on the basis of responses to environmental predictors, rather than presence-absence or abundance patterns. We first develop the model, including motivation, framework, and its application to multifarious data. We then discuss the role of prediction in GJAM. Finally we provide applications.

Model development

Consider species abundance data where adults are recorded on a continuous scale (e.g., basal area) and seedlings of the same or different species are recorded as discrete counts. We refer to these data types are *continuous abundance* (CA) and *discrete abundance* (DA), respectively. We wish to quantify their responses not only to environmental variables, but also their residual relationships to each other. For example, do they tend to covary, beyond what can be explained by environmental variables? Any transformations we might impose distort the scales and thus complicate interpretation. However, transforming data to different scales is not the only option. An alternative is available where discrete data are viewed as approximate (aggregated) versions of continuous data. This assumption is often implicit, as when counts (discrete) are used to model density (continuous) in the Poisson example above: y_{is} has the same scale as λ_{is}, but one is a discrete count, the other a continuous intensity.

An alternative means for integrating discrete and continuous data on the observed scales makes use of censoring, which affects weight of the observations and accommodates effort. For a specific example of sample weight that does not involve censoring, consider Poisson regression with a log link, which best predicts low values. The weight of an observation depends on its variance (e.g., Ver Hoef and Boveng 2007). Constant variance on the log scale places disproportionate weight on low values. There is nothing inherently 'correct' about this weighting, and it could be undesirable where low values are sporadic and noisy relative to large values, which could most important for fitting and prediction. Censoring affects the weight of an observation in a different way. Censoring extends a model for continuous variables across censored intervals. Survival analysis is a familiar example that can involve 'left-censored', 'interval-censored', or 'right-censored' observations. Continuous observations are uncensored. Discrete observations are censored and can depend on sample effort. Intensive effort in survival analysis, e.g., sampling daily rather than weekly or monthly, decreases the duration of censored intervals, decreases variance, and increases the weight of observations (Appendix S1). We learn most about mortality when all subjects die at times when sampling is frequent. We learn least when all subjects die within the same censored interval, which is most likely when intervals are long.

Censoring can be used with effort for an observation to combine continuous and discrete variables with appropriate weight. In composition data, effort is the total number of objects observed, e.g. the reads per observation in microbiome data. In census-count data, effort is determined by the size of the sample, search time, or both. It is comparable to the offset in generalized linear models (GLM). We discuss how these elements contribute to the model framework in the next section.

Model framework

Elements of the model are introduced first, followed immediately by a simple example demonstrating their relationships. We then consider applications to multiple data types.

A sample consists of n observations. Each observation i consists of two vectors, $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n$, where $\mathbf{x}_i$ is a vector of predictors $q = 1, \ldots, Q$, and $\mathbf{y}_i$ is a vector of responses y_{is}, with $s = 1, \ldots, S$. The combinations of continuous and discrete measurements in $\mathbf{y}_i$ are accommodated by locating each observed Y in two probability spaces, one continuous W and another discrete Z. In the previous example, basal area of trees has either zero or positive values. One way to model continuous data with zeros is the Tobit, introduced for economic data (Tobin 1958, Cameron and Trivedi 2005), but increasingly used in environmental applications, including agriculture (Bamire et al. 2002), precipitation (Sahu et al. 2010) and species distributions (Clark et al., 2014). In GJAM the two types of observations are identified by integer labels $z_{is} \in \{0, 1\}$. Positive values for y_{is} are assigned to a discrete interval $k = 1$. Zero values are assigned to the interval labeled $k = 0$ (Fig. 3a). In the Tobit model (and GJAM) fitting relies on a latent continuous variable w_{is}, which is known and equal to y_{is} when $y_{is} > 0$. When $y_{is} = 0$, the continuous variable w_{is} occupies the censored interval $z_{is} = 0$ and is known only to be negative.

We can extend this simple structure to accommodate each data type (Fig. 3) as follows. To generalize, a vector $\mathbf{w}_i \in \mathcal{R}^S$ locates $\mathbf{y}_i$ in continuous space. This continuous space allows for dependence between response variables with a covariance matrix. A second vector of integer values $\mathbf{z}_i \in \{0, \ldots, K-1\}^S$ locates $\mathbf{y}_i$ in discrete space. This discrete space allows for error in discrete observations, zero-inflation being the most common example. Each element of $\mathbf{z}_i$ assigns a corresponding element of $\mathbf{w}_i$ to an interval $z_{is} = k$. The number of intervals K can differ between observations and species, due to different levels of effort E_{is} and to different ways of observing different species. In other words K can have subscripts i, s, or both.

To connect continuous and discrete vectors there is a set of partition points $p_{is,k} \in \mathcal{P}$ that locate the continuous w_{is} within discrete intervals $z_{is} = k$. For now, assume that the partition does not differ between observations and species, $p_{is,k} = p_k$. Interval k is bounded by two points in the partition $(p_k, p_{k+1}]$. The intervals are contiguous and fully partition the real line $(-\infty, \infty)$. Unless there is zero-inflation, $k = 0$ has the partition $(p_0, p_1] = (-\infty, 0]$. The last interval is (p_K, ∞).

Finally, intervals are censored when observations are discrete; they are uncensored when observations are continuous. The set of censored intervals is $\mathcal{C}$, again, those intervals for which y_{is} is discrete, and w_{is} is unknown. Within uncensored intervals y_{is} is continuous and, thus, w_{is} is known.

For prediction, the model can be thought of like this: There is a vector of continuous responses $\mathbf{w}_i$ generated from mean vector $\boldsymbol{\mu}_i$ and covariance $\boldsymbol{\Sigma}$ (Fig. 4a). The partition $\mathbf{p}_{is}$ segments the continuous scale into intervals, some of which are censored and others not. A value of w_{is} that falls within a censored interval k generates observed $y_{is} = z_{is} = k$. A value of w_{is} that falls in an uncensored interval is assigned w_{is} (examples in Figure 3).

Of course, data present us with the inverse problem: the observed y_{is} are continuous or discrete, with known or unknown partition (Fig. 4b). The discrete class depicted for observed $y_{is} = 3$ in Figure 4b can correspond to a continuous w_{is} anywhere within the shaded interval on the W axis. Depending on how the data are observed, we must impute at least the elements of $n \times S$ matrix $\mathbf{W}$ that lie within censored intervals. Unknown elements of $\mathbf{Z}$ and $\mathcal{P}$ will also be imputed in order to estimate parameters (see below).

Before proceeding further, consider again the biomass example in Figure 1c for 98 tree species on forest inventory plots. Together, discrete zeros and continuous positive values define the $K = 2$ intervals, indexed as $k \in \{0, 1\}$. Because the partition is the same for all observations and species, all elements in the partition $\mathcal{P}$ can be represented by a length-$(K+1) = 3$ vector, $\mathbf{p} = (p_0, p_1, p_2) = (-\infty, 0, \infty)$. Because $k = 0$ is censored, and $k = 1$ is not, the set of censored intervals is a single value, $\mathcal{C} = \{0\}$. To get specific, if an observation vector for $S = 3$ species is $\mathbf{y}_i = (3.7, 0, 12.1)$, then $\mathbf{z}_i = (1, 0, 1)$, and $\mathbf{w}_i = (3.7, w_{i2} < 0, 12.1)$.

The advantage of this framework comes from the fact that modeling the contrasting data types commonly collected by ecologists requires no more than different combinations of known and unknown $\{W, Z, \mathcal{P}\}$. With variable effort and continuous y_{is} the w_{is} is known and equal to y_{is} (black lines in Figure 3). When y_{is} is discrete, the interval k is censored, w_{is} is imputed (grey lines in Figure 3), bounded by the two points in the partition $(p_{is,k}, p_{is,k+1}]$, with the i and s subscripts needed when there is differing effort between observations, species, or both. Discrete label z_{is} is imputed when there can be misclassification of discrete observations; zero inflation is an example (Fig. 3c). Zero inflation occurs when the recorded state is $y_{is} = 0$, and the true state is $z_{is} > 0$. Partition elements $p_{is,k}$ are imputed when the scale is unknown (e.g., ordinal data)(Fig. 3g).

The model combines each of the foregoing elements. The w_{is}, z_{is}, and $p_{is,k}$ differ for each data type and map to observations,

$$y_{is} = \begin{cases} w_{is} & \textit{continuous} \\ z_{is}, \quad w_{is} \in (p_{z_{is}}, p_{z_{is}+1}] & \textit{discrete} \end{cases} \tag{1}$$

where $p_{is,k} = p_{z_{is}}$. If there is no error in assignment of discrete intervals, then $z_{is} = k$ (the observed label is the true label), and the model for $\mathbf{w}_i$ is

$$\begin{aligned} \mathbf{w}_i | \mathbf{x}_i, \mathbf{y}_i &\sim MVN(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}) \times \prod_{s=1}^{S} \mathcal{I}_{is} \\ \mathcal{I}_{is} &= \prod_{k \in \mathcal{C}} I_{is,k}^{I(y_{is}=k)} (1 - I_{is,k})^{I(y_{is} \neq k)} \end{aligned} \tag{2}$$

where the indicator function $I(.)$ is equal to 1 when its argument is true and zero otherwise. The indicator

$$I_{is,k} = I(p_{is,k} < w_{is} < p_{is,k+1}) \tag{3}$$

means that w_{is} lies within the correct interval k. It applies only to the censored intervals, i.e., the set $\mathcal{C}$. The mean vector $\boldsymbol{\mu}_i = \mathbf{B}'\mathbf{x}_i$ contains the $Q \times S$ matrix of coefficients $\mathbf{B}$ and the length-Q design vector $\mathbf{x}_i$. $\boldsymbol{\Sigma}$ is a $S \times S$ covariance matrix. The partition depends on observation i if effort varies between observations (next section) and between responses s when they are observed on different scales. For ordinal data the partition is inferred (Fig. 3g). Eqn 2 is conditional on the discrete label $z_{is} = k$ being correct. The extension to incorrect z_{is}, including zero inflation, are given in Appendix S1.

The model accommodates the diversity of observations contained in field data. Extending the previous example, if large values are censored above a threshold U, e.g., a detector saturates or an observer does not count higher than $Y > U$, there will be $K = 3$ intervals with $K+1 = 4$ elements in the sample partition $\mathbf{p} = (-\infty, 0, U, \infty)$(Fig. 3b). Uncensored values fall in interval $z_{is} = 1$, defined by $0 < w_{is} < U$. An observation $\mathbf{y}_i$ can now take values on $[0, U]^S$, with point mass at both 0 and U. Between 0 and U values are continuous. In examples that follow each ecological attribute is accommodated by different combinations of known and unknown $\{W, Z, \mathcal{P}\}$, with a subset of intervals being censored, contained in the set $\mathcal{C}$.

Scale equivalence and the role of effort

Discrete data in ecology are often counts, which depend on the level of effort. That effort can differ between observations i and between species s within the same observation. In GJAM, effort enters through the partition $\mathcal{P}$, thus affecting the range of values for w_{is} in eqn (2). Where effort $E_{is} = 1$ the approach imposes no scale difference between $\mathbf{y}_i$ and $\mathbf{w}_i$, despite the fact that each response in $\mathbf{y}_i$ can have different scales. Before discussing how effort affects different types of observations we address the issue of scale.

Consider again a response vector that includes density of seedling counts and basal area of trees, corresponding to columns in matrix $\mathbf{B}$. Individual coefficients in this matrix $\beta_{q,s}$ describe the response of s to predictor q. They have scales of density/x_q for seedlings and of basal area/x_q for trees, where x_q is the dimension for predictor q. Likewise covariance $\mathbf{\Sigma}$ has scales of density × density (two seedling species), basal area × basal area (two tree species), and density × basal area (a tree and a seedling species). The coefficients and covariance have direct interpretation in terms of what is observed, because y_{is} is on the same scale as w_{is}. It can also be useful to compare species on the correlation scale, where $\mathbf{R}$ is the correlation matrix associated with $\mathbf{\Sigma}$ (Appendix S1).

Where there is no absolute scale, including presence-absence (PA), categorical (CAT), and ordinal count (OC) data, one is imposed. Observations recorded as $success/failure$ for presence-absence or *low/medium/high* for ordinal data are not absolute scales, but they have relative scales. We anchor the location of the first interval at zero and impose a unit-variance scale (Chib and Greenberg 1998). In other words, the correlation $\mathbf{R}$ is also the covariance $\mathbf{\Sigma}$.

Where effort $E_{is} \neq 1$ there is an effect on scale, allowing observations from different plot areas or composition counts to be included in the same analysis. For discrete counts, large plots must contribute more weight than small plots. Microbiome samples with high total reads must contribute more than those with few reads. To improve on current practice (e.g., McMurdle and Holmes 2014), effort should vary to account for the fact that observations with the most effort have the smallest variance and, thus, the largest effect on the fit.

GJAM achieves effort-based weighting through the partition. Where effort $E = 1$, the partition for discrete counts $0, 1, 2, \dots$ begin at $-\infty$, followed by midpoints between count values, $\mathbf{p} = (-\infty, 1/2, 3/2, \dots)$. For $z_{is} = k$ the interval is thus $(p_{i,k}, p_{i,k+1}] = (k - 1/2, k + 1/2]$. When effort varies between observations the partition shifts to the 'effort scale',

$$\mathbf{p}_{ik} = \left(\frac{k - 1/2}{E_i}, \frac{k + 1/2}{E_i} \right] \tag{4}$$

If observations are animals counted per hour, E_i can be search time. If observations are benthic organisms per sediment core, E_i can be core volume. If observations are seedlings per plot, then E_i can be the area of plot i. Because plots have different areas one might choose to model w_{is} on a 'per-area' scale (density) rather than a 'per-plot' scale. The upper portion of Table 1 compares two plots having counts that result in the same density of 100 trees per ha, but differ in plot area. The observation scale is counts per plot. The effort scale is area. The wide partition on a small 0.1-ha plot admits large variance around the observation of 10 trees per 0.1 ha plot; the partition width is 10 trees ha^{-1}. Conversely, a narrow partition on a larger 1.0-ha plot constrains density to a narrow interval (1 tree ha^{-1}) around the observed 100 trees per plot.

In microbiome data effort accommodates the differing reads per sample. The lower portion of Table 1 compares count composition data, where effort E_i is the total count for observation i, and w_{is} lies on the composition scale: when y_{is} is greater than zero and less than E_i, then $w_{is} \in (0, 1)$. Using the partition of eqn (4) the two observations that represent the fraction 0.10 in Table 1 with different effort (total reads in PCR data) are responsible for the declining predictive coefficient of variation in Figure 5*b*.

Censoring and effort combined are shown in Figure 5. A simulated example is shown in Figure 5a, where data are censored by the so-called 'octave scale', discrete observations recorded as $(0, 1, 2, 4, 8, \dots)$ (Preston 1948; Muller-Dombois and Ellensburg, 1974; Gauch 1982; Moore and Chapman, 1986, Jackson and Sullivan 2009). They are modeled with GJAM on this observation scale, allowing for increasing variance with increasing mean, a relationship that can be desirable, depending on application. Figure 5*b* exploits censoring to weight composition count data by effort per observation, in this case the number of reads from PCR data (see **Synthesis of microbiome data**).

Application to multifarious data

Attribute data differ only in terms of which of $W, Z, \mathcal{P}$ are observed versus imputed. Data types are summarized here and compiled in Table 2.

Continuous abundance (CA) data can be concentration, biomass, density, basal area, leaf area, cover, and so on. The previous section discusses how zeros and thresholds in continuous data are accommodated by censoring (Fig. $3a, b$). Where responses include zero GJAM provides an alternative to log transformation, which can place disproportionate weight on low values, does not allow zeros, and is not interpreted on the observation scale. The univariate counterpart of GJAM is a Tobit model. Previous application to multivariate data includes Clark et al. (2014).

Discrete abundance (DA) data arise from counts (Fig. $3e$). Count data are often not well described by standard distributions, such as the Poisson or the negative binomial, and perform poorly when zeros are common. The negative binomial can be more variable than the Poisson, but not less. When used for counts of multiple species, the multinomial distribution induces a negative covariance (e.g., Haslett et al. 2006, Paciorek and McLachlan 2009, deValpine and Harmon-Threatt 2013, Mandal et al. 2015). When the total count in the multinomial distribution is related to abundance a separate model is needed for this total (e.g., Royle 2004). By treating observed counts as a censored version of true abundance GJAM accommodates effort (Table 2), and parameters can be interpreted on the observation scale or the effort scale.

Presence-absence (PA) data include only two categories, $\{0, 1\}$(Fig. $3d$). The multivariate probit model of Chib and Greenberg (1998, see Pollock et al. 2014 for an ecological application) is a special case of GJAM for PA data, where both intervals are censored (Table 2). Because there is no scale, there is an imposed unit-variance scale.

Ordinal count (OC) data are collected where abundance must be evaluated rapidly, where precise measurements are difficult, or absolute scales are difficult to apply (Thuiller 2002). Because there is no absolute scale the partition must be inferred (Fig. $3g$). Consider the ordinal scale represented by categories with these labels: (absent, rare, intermediate, abundant). The sample partition is $\mathbf{p}_s = (-\infty, 0, p_{s,2}, p_{s,3}, \infty)$, where elements 2 and 3 are estimated (Fig. 3g). The zero anchors location, and unit variance imposes a scale. The model of Lawrence et al. (2008) is a special case for ordinal counts in GJAM (Appendix S1).

Composition data may be continuous fractions with a sum-to-one constraint (fractional composition) or discrete counts. Both have interpretation on the relative abundance $[0, 1]$ scale, and both require point mass at zero and one. Due to the sum-to-one (fractional composition) or sum-to-E_i (count composition) constraint, there is information on only $S - 1$ columns in $\mathbf{Y}$. *Composition-count* (CC) data are composition data reported as numbers of each species counted (Table 1). Composition counts are only meaningful in a relative sense; they provide no information on absolute abundance (Haslett et al. 2006, Paciorek and McLachlan 2009, de Valpine, and Harmon-Threatt 2013). The total count for a sample is the effort $E_i = \sum_s y_{is}$. Common examples include molecular sequence data (e.g., Lauber et al., 2009), paleoecology (Brewer et al. 2012, Haslett et al. 2006), and fungal assays (Saucedo-Garcia et al. 2014). In paleoecology total counts can differ widely between observations (www.neotomadb.org). The number of DNA sequence reads in microbiome data can range over orders of magnitude. A practice that is widespread in the microbiome community rarifies count data to achieve approximate equity between samples. This amounts to a massive manipulation of data that can throw away vast amounts of information. Alternative model-based approaches applied to counts are limited to single taxa (McMurdle and Holmes 2014). A multinomial model with second-stage covariance is not on the data scale. Moreover, dominance of zeros in microbiome data limits application of most approaches (Paulson et al. 2013, Li 2015).

GJAM accommodates the discrete observations and the underlying relative abundance scale. A sample count can take values $y_{is} \in \{0, 1, 2, \dots\}$, with E_i being the total count for sample i. The partition segments the $[0, 1]$ composition scale according to effort and allowing for zeros (Fig. 3f, Table 2) (Appendix S1). Small samples have wide bins and, thus, high variance and low weight (Fig. 5b).

Fractional composition (FC) data arise in many ways, examples including the fraction of a photoplot (Page et al. 2008) or remotely sensed image (Cohen et al. 2003) occupied by each species or cover type. It can be the fraction of leaves lost to different types of herbivory (Silfer et al. 2015) or stream or foliar chemistry (Ollinger et al. 2002). The correlations between responses are distorted when estimated on the multivariate logit scale (Fig. 2b). Still more problematic, the logit scale does not admit zeros, which are common in composition data (Adkinson 1986, Leininger et al. 2013). In a recent example Leininger et al. (2013) admit zeros by defining a reference response variable that does not include zeros. We could not obtain convergence with this model for data sets containing large numbers of zeros, particularly those where many observations are dominated by a single species. In GJAM a FC observation is represented in continuous space and censored at 0 (absent species) and 1 (monoculture)(Appendix S1).

A sample may have multiple composition groups. For example, $\mathbf{Y}$ may include both soil and endophytic microbiome data, each with its own total count (effort). Let G be the number of composition groups. If there are L_g response variables for a given FC or CC group g, then there are $L_g - 1$ non-redundant columns in $\mathbf{Y}$ for group g. A sample includes information on the total number of non-redundant columns, $S = \sum_g L_g - G$. A link function provides support over the real line for composition data, while providing estimates on the observation scale (Appendix S1).

Categorical data (CAT) describe unordered categories. If observation i refers to a sample plot, and the response s is a cover-type variable, then it might be assigned to one of several categories k, such as 'tidal flat', 'low marsh', or 'high marsh'. If it refers to a sample plant, and a response is growth habit, it might be assigned one of four categories 'herb', 'graminoid', 'shrub', or 'tree'. These are multinomial responses. Like composition data, a categorical response s occupies as many columns in $\mathbf{Y}$ as there are non-redundant levels $K_s - 1$, because the K_s columns sum to 1. The observed category is that having the largest value of $w_{is,k}$ for response s (Table 2). The model of Zhang et al. (2008) is a special case for the treatment of categorical responses in GJAM (Appendix S1).

These data types can be modeled jointly in the R package `gjam` at `https://cran.rstudio.com/web/packages/gjam/index.html`

Zero inflation

A zero-inflated model is used to boost the zero category for the purpose of better describing responses or to allow both for an underlying process that admits zero (e.g., a population cannot persist at a site) and for observed zero when the underlying process is not zero (the population can persist, but is not detected). The simplest approach uses the effort-based partition in eqn 4 to expand the $k = 0$ category,

$$\mathbf{p}_{i,0} = \left(-\infty, \frac{1}{2E_i}\right] \tag{5}$$

Note that the second value is greater than zero, but it approaches zero with increasing effort– effort decreases the probability of missing the species. The second approach to zero inflation is to model the miss-classification of the discrete state (Appendix S1). In this case the label z_{is} must be estimated together with w_{is} and parameters (Fig. 3*c*).

Model fitting

Model fitting entails simultaneous inference on parameters $(\mathbf{B}, \boldsymbol{\Sigma})$, together with latent states in $\mathbf{W}, \mathbf{Z}$, and any that are unknown in the partition $\mathcal{P}$, depending on each observation type in the sample (Table 2). Posterior simulation is done with Gibbs sampling in the R package `gjam` (Appendix S1), written in R (R Development Core Team 2013) and C++ (Clark 2016). Prior distributions are discussed in the Appendix S1. Latent variables are sampled subject to the partition (eqns $2, 4$). Regression coefficients are sampled from the matrix normal distribution with a non-informative prior. The covariance matrix is sampled from the inverse Wishart distribution where regression coefficients are marginalized. Where the scale is unknown (presence-absence, ordinal, nominal) parameter expansion is used to sample on the correlation scale. For ordinal data the partition is sampled (Lawrence et al. 2008). Zero inflation involves an additional step to sample the discrete label z_{is} when $y_{is} = 0$ (Appendix S1).

Roles for prediction

The covariance $\boldsymbol{\Sigma}$ plays a prominent role in predicting relationships between species. Matrix $\boldsymbol{\Sigma}$ is the covariance between species after removing relationships explained by the mean structure of the model, $\boldsymbol{\mu}_i$ in eqn 2. On the one hand, it is important to demonstrate that $\boldsymbol{\Sigma}$ is identified in the model, as we do with examples that follow. It is equally important to recognize that a model that explains much of the variation in data has high signal-to-noise, $|\mathbf{B}'\mathbf{x}_i| \gg \sqrt{diag(\boldsymbol{\Sigma})}$. In other words, we seek to concentrate variation in $\boldsymbol{\mu} = \mathbf{B}'\mathbf{X}$. When this goal is achieved diagonal elements of $\boldsymbol{\Sigma}$ are small, and off-diagonals are indistinguishable from zero. Non-zero off-diagonals mean that species still have information to convey on the abundance of others, after accounting for $\boldsymbol{\mu}$. Given $\boldsymbol{\mu}$, marginal independence between species s and s' means that $\boldsymbol{\Sigma}_{s,s'}$ does not differ from zero. Potentially of greater interest, conditional independence means that $\boldsymbol{\Sigma}^{-1}_{s,s'}$ does not differ from zero (Rajaratnam et al. 2015). Conditional independence means that there is no evidence for a direct relationship between two species. Alternatively, non-zero $\boldsymbol{\Sigma}^{-1}_{s,s'}$ finds evidence for a relationship between species that does not come from their mutual relationships to other species or from $\boldsymbol{\mu}$. The applications of prediction that follow involve estimates of $\boldsymbol{\Sigma}$ and the role they play in i) missing data imputation, ii) variable selection, iii) sensitivity analysis, and iv) species clustering.

Characterizing communities

The long tradition in ecology of defining communities is primarily based on correlation or distance matrices evaluated for empirical data (Gauch 1992, ter Braak and Prentice 1988). Joint models provide opportunity to examine community structure probabilistically on the basis of environmental responses, with full uncertainty. The $Q \times S$ matrix $\mathbf{B}$ contains relationships of each species to the environment–the 'signal'–but not to each another. A predictive approach can translate $\mathbf{B}'\mathbf{X}$ to an $S \times S$ covariance among species. This translation requires a distribution for a vector of predictors $\tilde{\mathbf{x}}$; the observed $\mathbf{x}_i$ are fixed ($\mathbf{x}_i$ is deterministic in the model), but we can assign a distribution to $\tilde{\mathbf{x}}$ as a scenario, justifying the approach (Appendix S1). Consider a distribution of centered input variables having structure like that of observations,

$$\tilde{\mathbf{x}} \sim MVN(0, \mathbf{V}) \tag{6}$$

where $\mathbf{V}$ is a covariance matrix for $\tilde{\mathbf{x}}$. Marginalizing $\tilde{\mathbf{x}}$ contributes the environmental component of variation in response $\tilde{\mathbf{y}}$,

$$\mathbf{E} = \mathbf{B}'\mathbf{V}\mathbf{B} \tag{7}$$

Eqn 7 has the dimensions of a species covariance matrix ($Y_s \times Y_{s'}$), and it has a corresponding correlation matrix $\mathbf{R}^{\mathbf{E}}$. It is not the correlation matrix reported by Pollock et al. (2014). When $S > Q$ (all examples given here) $\mathbf{E}$ is not full rank and thus does not have an inverse. We can evaluate a Moore-Penrose pseudoinverse. Matrix $\mathbf{E}$ summarizes species similarities in terms of their response to an environment $\tilde{\mathbf{x}}$. Similar species have similar columns in $\mathbf{B}$. Those similarities and differences are amplified for predictors $\tilde{\mathbf{x}}$ with large variance. Conversely, species differences in $\mathbf{B}$ do not matter for variables in $\mathbf{X}$ that do not vary. The covariance in predictors could come from observed data, i.e., the variance of $\mathbf{X}$, in eqn 6. It could represent a subset of the data, e.g., that for a particular region. It could be a scenario for future conditions.

Sensitivity analysis

In univariate models each element of vector $\mathbf{B}$ is a sensitivity coefficient, the effect of one predictor in $\mathbf{X}$ on one response in $\mathbf{y}$. Coefficients can be compared to evaluate the importance of $Q - 1$ inputs in $\mathbf{x}$ (omitting the intercept). In multivariate models coefficients in the $Q \times S$ matrix $\mathbf{B}$ do not quantify the overall importance of predictors. The S coefficients associated with each predictor cannot be added together or averaged. Inverse prediction integrates all S responses in $\mathbf{y}_i$, thus reducing sensitivity analysis from $S \times (Q - 1)$ coefficients to $Q - 1$

coefficients, i.e., one per predictor variable (Clark et al. 2011, 2013). For a model that is linear in $\mathbf{X}$ the predictive distribution from eqn 6 is $\tilde{\mathbf{x}} \sim MVN(\mathbf{m}, \mathbf{V})$, where $\mathbf{m}$ is mean vector, $\mathbf{V}^{-1} = \mathbf{F} + \mathbf{U}^{-1}$ is the covariance, and $\mathbf{U}$ is the prior covariance matrix for $\mathbf{x}$. The quantity

$$\mathbf{F} = \mathbf{B}\mathbf{\Sigma}^{-1}\mathbf{B}' \tag{8}$$

is the 'information' contributed by the fitted coefficients. These predictions can be compared using prediction scores (Gneiting and Raftery 2007) against the true values of $\mathbf{x}$ (Clark et al. 2013; 2014). An accurate and not-overconfident prediction has a high prediction score. Brynjarsdottir and Gelfand (2015) suggest that the diagonal be used as a sensitivity coefficient,

$$\mathbf{f} = diag(\mathbf{F}) \tag{9}$$

In both cases the importance of each covariates in $\mathbf{X}$ is summarized by a single value f_q, integrating all information in the model.

Missing data and model selection

Species abundance data sets can be large and heterogeneous, often having missing values. The predictive distributions for $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{x}}$ allow imputation as part of Gibbs sampling (Appendix S1). Missing values become part of the posterior distribution.

Prediction can also be used for model selection. Model selection can be based on parameter space (e.g., AIC, DIC) or predictive space (Gelfand and Ghosh 1998; Hooten and Hobbs 2015; Dawid and Musio 2015). Advantages of the latter include the fact that the interpretation of parameters changes with the model, but predictive space does not; it makes sense to criticize models in terms of their capacity to predict the data (in- and out-of-sample). We use DIC and the Gneiting and Raftery (2007) prediction score.

Model summary

In summary, for data all of one type, GJAM generalizes existing multivariate (MV) models, including the MV probit (Chib and Greenberg 1998), MV Tobit (Clark et al. 2014), MV ordinal (Lawrence et al. 2008), and MV nominal (Zhang et al. 2008) models. It extends to new data types (discrete counts, composition), accommodating their differences through a partition that links continuous and discrete states and effort. Each of these methods can be viewed as special cases of eqn 2 (Table 2). Each data type involves a coefficient matrix $\mathbf{B}$ and a covariance matrix $\mathbf{\Sigma}$. Depending on a partition $\mathcal{P}$, which incorporates effort E, parameters generate continuous W and, when it is unknown, discrete Z. In the case of ordinal data the partition is also estimated. For presence-absence, ordinal, and categorical data $\mathbf{\Sigma}$ is a correlation matrix $\mathbf{R}$.

So one size fits all, but the framework can go further. The same model applies when the different data types are *modeled together*. In GJAM, the partition and selective use of parameter expansion allows modeling with eqn (2), where each column of $\mathbf{Y}$ can be a different data type. In the diagnostics and applications that follow we show how it applies to combined data.

Diagnostics

Simulated data

To determine if GJAM recovers true parameter values and can predict data we conducted simulations. Simulation steps include 1) specify partition $\mathcal{P}$ for different data types, 2) generate random parameter values $(\mathbf{B}, \mathbf{\Sigma})$ and design $\mathbf{X}$, 3) draw a sample $\mathbf{W}$, and 4) partition $\mathbf{W}$ with $\mathcal{P}$ to obtain $\mathbf{Z}$ and $\mathbf{Y}$ (eqn 2). Posterior distributions were simulated to confirm parameter identifiability and data prediction.

Figures 6, 7 illustrate joint modeling with a mixture of attributes that includes ordinal counts (e.g., host or plot condition, qualitative assessments), presence-absence (e.g., potential pathogens, predators, herbivores), continuous abundance (e.g., basal area, biomass, nutrient concentration), discrete abundance (e.g., number of seedlings), count composition (e.g., microbiome data), and continuous without censoring. Coefficients for all data types are estimated jointly (Fig. 6*a*), including the correlation matrix (Fig. 6*b*). The partition matrix for ordinal data is recovered (Fig. 6*c*). The fitted model predicts all data types well, despite contrasting scales (Fig. 7). Predictions are least accurate where there are small numbers of observations, shown as histograms below predictions in Figure 7. Extensive simulation studies were used to determine that the model predicts disparate species groups and attributes, each informing the others in ways that can contribute to prediction.

To determine the effect of collapsing abundance data into presence-absence, we compared estimates for simulated abundance data fitted in two ways, one as abundance and another as presence-absence. We found that excellent parameter recovery on the abundance scale (Fig. 8, left) does not translate to the presence-absence analysis, particularly the correlation matrix (Fig. 8, right). Even presence-absence is predicted better by the abundance model than by the presence-absence model (Fig. 8, lower panels). Furthermore, presence-absence models cannot admit any species that are present at all sites, i.e., *the most abundant species.* Thus, GJAM allows us to evaluate the consequences of discarding abundance information and shows that effects can be substantial.

GLM comparisons

We compared GJAM with current practice based on GLMs. Comparisons with simulated data have the advantage that 'true' parameter values are available from simulation, but they should be further checked with real data, which, of course, do not have a 'correct' model. We wanted to know if the Gaussian first-stage model was unrealistic and thus might perform poorly in comparison to standard link functions in GLMs.

Figure 9*b* compares a standard GLM model (Poisson likelihood with log link) with GJAM for stem counts on FIA data (data used in the Section **Forest inventory in eastern North America**), using the same predictors in $\mathbf{X}$. The GJAM root mean square prediction error (rmspe) is half that of the GLM. The modal predictions for GJAMs are consistently closer to the data than for the GLM. The downward bias in the Poisson model is pronounced at high values, because the log link emphasizes the lowest values. GJAM does not differentially weight observations by abundance alone and is much more accurate than the GLM at high values, which, again, might often be of most interest. Thus, the linear link and Gaussian assumptions in GJAM perform better, not worse, than the standard model. It has the further appeal that parameter estimates are on the same scale as the observations and thus have transparent interpretation.

Differences are still more striking for the Bernoulli example in Figure 9a, where the rmspe for the GLM is 37-fold larger than GJAM. Both models involve the probit, and they have the same mean structure. The models differ in that GJAM jointly models host status (Fig. 1a) together with its endophytic microbiome (Fig. 1b), composition data. In other words, GJAM synthesizes multiple data types, while still offering superior prediction for each individually.

In summary, although the Gaussian assumption of GJAM could be criticized as being unrealistic for real data, we show that it performs better than standard models widely used in ecology. To determine if performance is improved by the generalizing the Gaussian to asymmetric distributions we have implemented the skew-normal (Azzalini 2005), including for composition data, and find negligible benefit despite substantially greater complexity (Taylor-Rodriquez et al., in prep).

Applications

Forest inventory in eastern North America

Just as the environment controls distributions of species (Cowles 1911, Sinclair et al. 2010), the biodiversity of a site might hold clues to the environment. The promise that vegetation might reveal underlying environmental conditions has motivated its use for water and mineral prospecting (Brooks 1979), disease risk (Robinson et al. 1997), climate reconstruction (Brewer et al. 2012), and conservation (Larson et al. 2004, Nichols and Williams 2006). But individual species or aggregate vegetation characteristics (e.g., remote sensing) tend to be limited in their indicator value (Ellenberg 1982, Dufrene, and Legendre. 1997, Cannon 1971, Brooks 1979, Gmez-Girldez et al. 2014). For example, most soil types and terrain offer only slight advantages for some species over others, and most species still occupy a broad range of sites (Whittaker 1978). GJAM provides a first opportunity to predict site conditions probabilistically, without need for indicator species, through inverse prediction from the full (joint) model.

This example uses USDA Forest Inventory and Analysis (FIA) data to combine species-level data with plot-level data. We demonstrate application with variables at these different scales, including inverse predictive of the environment. Responses are plot-level foliar N and P, both continuous responses as community-weighted mean values (Clark 2016), together with biomass of tree 98 species that occurred on at least 50 plots, all continuous abundance with point mass at zero; there are a total of $S = 2 + 98 = 100$ responses. FIA data come from 0.0672-ha plots established at a density of 1 per 2428 ha (Bechtold and Patterson 2005, Woudenberg et al. 2010, USDA 2012). All trees $>$ 12.7 cm in diameter are counted and measured. Individual plots are so small that each species is represented by, at most, a few individuals, and many species present in an area will be absent simply due to small plot size. For this reason analyses are often based on aggregate plots (Iverson and Prasad 1998, Zhu et al. 2014, Clark et al. 2014). For this illustration we aggregate 19,568 FIA plots into 1617 one-ha plots, a $k-$means clustering using covariates (Schliep et al. 2015). In other words, plots are similar in covariate space. Most observations (72%) are zero. Predictors in the model include temperature, moisture, local terrain (slope, aspect), and soil type. Slope and aspect are represented by a length-3 vector specified in the caption of Figure 11. Predictors have low correlation with one another and low variance inflation factors (Appendix S1). Computation makes use of the dimension reduction algorithm of Taylor-Rodriquez et al. (2016), although a data set of this size does not require it (Clark et al. 2014).

We first determined that the model predicted the responses (Fig. 10), including the overall plot richness, which was not actually fitted with the model (Fig. 10c). We include this because SDMs over-predict richness (Guisan and Rahbek 2011, Clark et al. 2014). Accurate but wide predictive intervals for the continuous foliar traits reflect the that fact that these are plot-level variables, contributed by species with a broad range of foliar N and P values (Fig. 10a). Continuous abundance predictions for tree biomass are broad for non-zero observations, because most are rare (histogram at the base of Figure 10b). Likewise, the species richness predictions are poor for the most- and least-diverse sites, because these sites are rare (Fig. 10c), but are otherwise accurate.

Soil types and slope emerge as the most important predictors in the model (Fig. 11). They account for the largest effects on individual species (Figure 11, right). The predictive distributions for overall sensitivity $\hat{\mathbf{F}}$ (eqn 8) are highest for two soil types, the ultisols that dominate the eastern Piedmont and the mollisols most prevalent in the Upper Midwest (Fig. 11, left). Despite the strong effect of slope (u_1), aspect effects (u_2, u_3) are weak for all species (Fig. 11, right).

Despite the fact that individual predictors show that slope effects are large for few species, and aspect effects are weak for all species (Fig. 11), the full model allows precise inverse prediction of the local environment. Taking aspect as an example, effects are evident in only a small subset of species, with mesic species biased toward the NE (Fig. 12). Even for the most responsive species, effects are subtle, less than 5 m^2 ha^{-1} basal area on 20° slopes. Despite weak site effects for species individually inverse prediction provides precise predictive capacity not only for regional temperature (Fig. 13a), but also for local habitat, including moisture, slope, and aspect (Fig. 13b, c, d). By exploiting information for all species together inverse prediction identifies habitats where no individual species could. These results indicate that the species modeled jointly can be used to predict local site conditions, despite the fact that individual species cannot.

The model further indicates that structure in abundance data does not provide an accurate representation of environmental responses in the model. Standard methods for identifying structure on ecological communities build from co-occurrence or abundance data. Figure 14a shows the species × species correlation matrix, a starting point or close relative of similarity matrices used for many clustering and ordination methods (e.g., Oksanen 2008). The order of species in Figure 14a follows a cluster analysis to highlight similarities among species. A complete-linkage algorithm was used in the R package stats::hclust (R Core Team). This and other clustering algorithms we applied found only weak pattern in the data. With the exception of few 'red' combinations in Figure 14*a*, correlations are almost entirely in the range from -0.2 to 0.2. The response matrix $\hat{\mathbf{E}}$ in Figure 14*b* from eqn 7 is assembled in the same order as Figure 14*a*. If the variation in field data was explained by the model, then patterns in the two should be similar. They are not; the dense mixture of high positive (red) and negative (blue) values in Figure 14*b* means that the structure in field data is quite different from the structure of responses.

However, when we reorganize $\hat{\mathbf{E}}$ according to its own structure there are clear species assemblages (Fig. 14*c*). The strong contrasts in colors, clearly organized in species groups, shows that structure in the *response* is dramatic and not well-captured by the tendency to co-occur.

Synthesis of microbiome data

Synthesis of data collected and analyzed by different methods and for different purposes is a goal of microbiome research (Gilbert et al. 2014). Synthesis is challenging, due to the size of sequence data (Lauber et al. 2009), over-representation of zeros, variable effort of composition data, and the fact that few studies collect ancillary data needed for model fitting and prediction. The large number of operational taxonic units (OTUs) generated by sequence methods poses a 'big-S, small-n' problem; S can be orders of magnitude larger than n. Dimension reduction schemes seek to zero out elements of $\mathbf{B}$, $\mathbf{\Sigma}$, or $\mathbf{\Sigma}^{-1}$ or to reduce the rank of $\mathbf{B}'\mathbf{X}$ or $\mathbf{\Sigma}$ (e.g., Pati et al. 2014; Rajaratnam et al. 2015; Goh et al, 2015). Thus far, microbiome data have been evaluated primarily with descriptive techniques, to identify groups of taxa that could be related in where they occur and how they respond to the environment. The inconsistency in covariates means that a given predictor variable is likely to be absent for many samples. Finally, the sampling effort varies over orders of magnitude, the number of reads per sample (Fig. 1*b*). This variation has led to the practice of rarifying samples down to some common sum, thus discarding the bulk of the information (McMurdle and Holmes 2014). We focus on dimension reduction for the GJAM in a separate study (Taylor-Rodriquez et al, in revision) focusing here on the more fundamental question of potential for model-based analysis of microbiome data.

Data for this example come from the Earth Microbiome Project (EMP) global soils database, a project initiated to standardize molecular phylogenetic approaches across datasets to facilitate comparisons within and between studies (Gilbert et al. 2014). This composite data set provides no common predictors other than latitude and a habitat variable. The second most frequent variable is pH, which is available for only 245 (50% of) studies. These challenges are common for data compilations. The example provides opportunity to examine if effective inference for such combined data sets can be done despite the high degree of data imputation, for median-zero data, and few covariates.

To illustrate GJAM application to composition data we extracted all OTUs that occur in at least 350 samples. Typical of molecular phylogenetic data, observations are dominated by soil bacteria, primarily Acidobacteria and Proteobacteria. Estimates integrate the heterogeneous effort represented by samples that range over four orders of magnitude in total reads (Fig. 15). GJAM imputes missing values, but we anticipate that massive missingness will degrade the fit. The effect of effort comes through the weight contributed by samples, those with least effort having the highest variance (Fig. 5*b*) and thus the weakest contribution. Predicted abundance is imprecise (not shown), reflecting tremendous scatter in the data, primarily zeros, few predictors to include in the model (pH, latitude), and massive imputation of input variables (50% for pH, and two latitude values). Still, sensitivity estimates show clear differences between inputs, including a stronger effect of latitude than pH. They further indicate some capacity to inverse-predict pH and local habitat, but not latitude, from the fitted model (Fig. 16). Clear structure in the **E** matrix is indicated by red blocks at left in Figure 17. On the standardized scale pH and latitude have little impact in comparison (right side of Figure 17).

The fact that half of all pH data had to be estimated (blue dots in Fig. 16*b*) together with coefficients suggests that improvement will come simply from greater availability of predictor variables. Even with these limitations, GJAM shows that microbiome data can be used to predict habitat (Fig. 16*c*), if not the reverse. These estimates highlight the importance of some standard set of predictors deemed important for the microbiome that would be encouraged from all investigators. We are now engaged in an extensive analysis of individual data sets where there are many inputs.

Discussion

The GJAM framework accommodates the median-zero, multivariate, multifarious nature of attribute data with an explicit connection between discrete and continuous observations on all species simultaneously (Fig. 3). The framework extends joint species distribution modeling to generalized joint attribute modeling (GJAM). Avoiding the transformation and rescaling that is needed with alternative methods facilitates interpretation of correlation structure on the observation scales. Advantages resolve some important challenges for species distribution models (SDMs) and joint species distribution models (JSDMs), including those that consider abundance (Latimer et al. 2009; Thorson et al. 2015).

A first advantage is accurate prediction. Recent studies note the challenges of prediction from species distribution models (Baselga and Araujo 2010, Guisan and Rahbek 2011, Clark et al. 2014). The accurate predictions for multifarious data with GJAM relies on proper treatment of continuous and discrete data, including overwhelming zeros. We verify parameter recovery and predictive performance in simulation (Fig. $6, 7, 8$). We demonstrate some advantages over standard methods for probabilistic prediction (Fig. 9). Although GJAM avoids the scale distortion that comes with a non-linear link function it predicts data better, not worst, than standard GLMs (Fig. 9). Unlike algorithmic-based methods, such as regression trees, it provides sensitivities to all inputs and species covariance, with full uncertainty.

The capacity to infer and interpret relationships between species on the observation scale avoids the distorted correlations that result from fitting hierarchical models with link functions (Fig. 2). For data that lack an absolute scale, presence-absence, nominal, and ordinal, the imposed unit-variance scale still permits parameter recovery and accurate prediction, including their relationships with other species that do have an observation scale (Fig. $6, 7$). These relationships range from a simple tendency to co-occur (presence-absence data), to possess attributes that co-occur (categorical data), to co-occur within similar ordinal categories, and to co-occur at similar absolute abundances (other data types).

Inverse prediction (IP) is especially valuable in the joint setting, not only for missing data imputation, but also for extracting the role of input variables (Fig. $13, 16$). IP provides detailed insight on the environment by combining the information in all species and the model. Although microbiome diversity is not well predicted by the environment, results show promise that the environment can be inversely predicted from the microbiome (Fig. 16c). Although 'indicator species' are rarely available for important environmental variables, the full community can provide precise insight (Fig. 13). For sensitivity analysis IP reduces the contributions from 10^3 parameter values in $\mathbf{B}$ and $\mathbf{\Sigma}$ to $Q-1$ sensitivity coefficients (Fig. 11).

The question of how many species to model requires a few technical remarks. We do not report here on dimension reduction methods for the GJAM, but it accommodates them (Taylor-Rodriquez et al., in revision). Most ecological data sets do not involve thousands or even dozens of species. For those that do include many species, a hard limit on the total number of species that can be modeled depends on n, just as a hard limit on the number of predictors in $\mathbf{B}$ (in absence of dimension reduction) cannot exceed n. The covariance matrix $\mathbf{\Sigma}$ must be full rank to allow inversion and model fitting. A prior distribution can rescue an otherwise non-invertible $\mathbf{\Sigma}$, but then the prior dominates. By marginalizing regression coefficients in our sampling of $\mathbf{\Sigma}$ (Appendix S1) we avoid high sensitivity to a prior at the cost of requiring that $\mathbf{\Sigma}$ is full rank. Long before a hard limit on number of species is reached we expect a degraded fit. Our applications show GJAM working well for 10^2 species. Given that microbiome data are dominated by zeros (Fig. 15), many applications may still work with subsets or aggregations of sequence data. As mentioned above, productive developments can focus on rank reduction, in which case many more species can be included (Taylor-Rodriquez et al., 2016).

In conclusion, GJAM provides new flexibility for inference and prediction from ecological data. GJAM aligns the scales for observations of many types and fits the model on observation scales.

Acknowledgements

The project was funded by the Macrosystems Biology, EAGER, and LTER programs of the National Science Foundation (NSF-EF-1137364, NSF-EF-1550911). For discussion and review we thank Bene Bachelet, Janet Franklin, Alan Gelfand, Brian Inouye, Kimberley Kauffman, Andrew Latimer, Daniel Taylor-Rodrigues, Cajo ter Braak, Bradley Tomasek, the participants of the NSF's SAMSI program on Ecological Statistics, and three anonymous reviewers.

Finally, we remember some good years with Diana Nemergut, the stimulating collaborations, generosity, and warm friendship.

Literature citations

–

Ackerly, D. D. and W. K. Cornwell. 2007. A trait-based approach to community assembly: partitioning of species trait values into within- and among-community components. Ecology Letters 10, 135?145.

Aitchison, J. 1986. The Statistical Analysis of Compositional Data, Chapman and Hall, New York.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association, 88, 669?679.

Araujo, M.B. and Luoto, M. 2007. The importance of biotic interactions for modelling species distributions under climate change. Global Ecology and Biogeography 16,743?753.

Azzalini, A. 2005. The skew-normal distribution and related multivariate families, Scandinavian Journal of Statistics, 32, 159-188.

Bamire A. S., Fabiyi Y. L., Manyong V. M. 2002. Adoption pattern of fertiliser technology among farmers in the ecological zones of south-western Nigeria: a Tobit analysis. Australian Journal of Agricultural Research 53, 901?910.

Baselga, A. and Araujo, M.B. 2010. Do community models fail to project community variation effectively? Journal of Biogeography. 37: 1842-1850.

Benito, B.M., L. Cayuela, and F. S. Albuquerque. 2013. The impact of modelling choices in the predictive performance of richness maps derived from species-distribution models: guidelines to build better diversity models. Methods in Ecology and Evolution, 4, 327-335.

Booth, T.H., H. A. Nix, J. R. Busby and M. F. Hutchinson (2014) BIOCLIM: the first species distribution modelling (SDM) package, its early applications and relevance to most current MaxEnt studies. Divers. Distrib. 20, 1?9.

Botkin, D. B., H. Saxe, M. B. Araujo, R. Betts, R. H. W. Bradshaw, T. Cedhagen, P. Chesson, T. P. Dawson, J. R. Etterson, D. P. Faith, S. Ferrier, A. Guisan, A. S. Hansen, D. W. Hilbert, C. Loehle, C. Margules, M. New, M. J. Sobel, and D. R. B. Stockwell. 2007. Forecasting the effects of global warming on biodiversity. Bioscience, 57, 227-236.

Brewer S, Jackson S.T. and Williams, J.W. (2012) Paleoecoinformatics: Applying geohistorical data to ecological questions. Trends in Ecology and Evolution, 27, 104-112.

Brooks, R.R. 1979. Indicator plants for mineral prospecting ? a critique. Journal of Geochemical Exploration. 12: 67-78.

Brynjarsdottir, J. and A.E. Gelfand. 2014. Collective sensitivity analysis for ecological regression models with multivariate response. Journal of Biological, Environmental, and Agricultural Statistics, 19, 481-502.

Calabrese, J. M., Certain, G., Kraan, C. and Dormann, C. F. (2014). Stacking species distribution models and adjusting bias by linking them to macroecological models. Global Ecol.Biogeogr., 23, 99-112.

Cameron, A. C. and P. K. Trivedi. 1998. Regression Analysis of Count Data. New York, Cambridge University Press.

Cameron, A. C. and P. K. Trivedi. 2005. Microeconometrics: Methods and Applications, Cambridge University Press, New York.

Cannon, H.L. 1971. The use of plant indicators in ground water surveys, geologic mapping, and mineral prospecting. Taxon 20: 227-256.

Chakraborty, A., Gelfand, A.E., J.A. Silander, Jr., A.M. Latimer, and A.M. Wilson. (2010) Modeling large scale species abundance through latent spatial processes. Annals of Applied

Statistics, 4, 1403-1429.

Chib, S. (1998). Analysis of multivariate probit models. Biometrika, 85:347

Chib, S., and Greenberg, E. (1998) Analysis of multivariate probit models, Biometrika, 85, 347-361.

Clark, J.S. 2016a. gjam: Generalized Joint Attribute Modeling . Comprehensive R Archive Network (CRAN). `https://cran.rstudio.com/web/packages/gjam/index.html`

Clark, J.S. 2016b. Why species tell us more about traits than traits tell us about species: Predictive models. Ecology, 97, 1979-1993.

Clark, J.S., D.M. Bell, M.H. Hersh, M. Kwit, E. Moran, C. Salk, A. Stine, D. Valle, and K. Zhu. (2011) Individual-scale variation, species-scale differences: inference needed to understand diversity. Ecology Letters 14, 1273-1287.

Clark, J.S., D. M Bell, M. Kwit, A. Powell, and K. Zhu. (2013) Dynamic inverse prediction and sensitivity analysis with high-dimensional responses: application to climate-change vulnerability of biodiversity. Journal of Biological, Environmental, and Agricultural Statistics, 18, 376-404.

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

Cohen, W.B., T.K. Maiersperger, Z. Yang, S.T. Gower, D.P. Turner, W.D. Ritts, M. Berterretche, S.W. Running. 2003. Comparisons of land cover and LAI estimates derived from ETM+ and MODIS for four sites in North America: a quality assessment of 2000/2001 provisional MODIS products. Remote Sens. Environ. 88, 233?255.

Cogan R.D. and Diefenbach D.R. (1998) Effect of undercounting and model selection on a sightability-adjustment estimator for elk. Journal of Wildlife Management, 62, 269?279.

Collins, P. C., Kennedy, R. and Van Dover, C. L. (2012) A biological survey method applied to seafloor massive sulphides (SMS) with contagiously distributed hydrothermal-vent fauna. Marine Ecology Progress Series, 452, 89-107.

Cowles, H. C. (1911) The causes of vegetational cycles. Annals of the Association of American Geographers, 1: 3-20.

Cunningham, R.B., and D. B. Lindenmayer. (2008) Modeling count data of rare species: some statistical issues. Ecology 86, 1135-1142.

Dawid, A.P. and M. Musio. (2015) Bayesian model selection based on proper scoring rules. Bayesian Analysis 10, 479-499.

de Valpine, P. and A. N. Harmon-Threatt 2013. General models for resource use or other compositional count data using the Dirichlet-multinomial distribution. Ecology 94:2678?2687.

Daz, S., Lavorel, S., de Bello, F., Qutier, F., Grigulis, K., and Robson, T. M. (2007) Incorporating plant functional diversity effects in ecosystem service assessments. Proceedings of the National Academy of Sciences, 104, 20684?20689.

Dufrene, M., and P. Legendre. 1997. Species assemblages and indicator species: The need for a flexible asymmetrical approach. Ecological Monographs 67:345-366.

Dunstan, P. K., Foster, S. D., Hui, F. K., and Warton, D. I. (2013). Finite mixture of regression modelling for high-dimensional count and biomass data in Ecology. Journal of Agricultural, Biological and Environmental Statistics, 18:357?375.

Elith J, et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29: 129-151.

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.

Ellenberg H. 1982. Vegetation Mitteleleuropas mit den Alpen in okologischer Sicht. Verlag Eugen Ulmer, Stuttgart, Germany, 989 pages.

Ferrier, S., Drielsma, M., Manion, G., andWatson, G. (2002). Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. II. Community-level modelling. Biodiversity and Conservation, 11:2309?2338.

Ferrier, S., Manion, G., Elith, J., and Richardson, K. (2007). Using generalized dissimilarity modeling to analyse and predict patterns of beta diversity in regional biodiversity assessment. Diversity and Distributions, 13:252-264.

Finley, A.O., S. Banerjee, and R.E. McRoberts. (2009) Hierarchical spatial models for predicting tree species assemblages across large domains. Annals of Applied Statistics, 3, 1052?1079.

Garnier, E., Cortez, J., Bills, G., Navas, M.-L., Roumet, C., Debussche, M., Laurent, G., Blanchard, A., Aubry, D., Bellmann, A., Neill, C. and Toussaint, J.-P. (2004) Plant functional markers capture ecosystem properties during secondary succession. Ecology, 85, 2630?2637.

Gauch, H.G. (1982) Multivariate Analysis in Community Ecology. Cambridge University Press, Cambridge, England.

Gelfand, A. E., and S. K. Ghosh. (1998) Model choice: a minimum posterior predictive loss approach. Biometrika, 85, 1?13.

Gelfand, A.E., J. A. Silander, S. Wu, A. Latimer, P. O. Lewis, A. G. Rebelo and M. Holder. (2006) Explaining species distribution patterns through hierarchical modeling. Bayesian Analysis, 1, 41-92.

Ghosh, S., A.E. Gelfand, K. Zhu, and J.S. Clark. (2012) The k-ZIG: flexible modeling for zero-inflated counts. Biometrics, 68, 878-85.

Gilbert, J.A., J. K. Jansson, and R. Knight. (2014) The Earth Microbiome project: successes and aspirations. BMC Biology 2014, 12:69 doi:10.1186/s12915-014-0069-1

Goh, G., K. Chen, and D.K. Dey. (2015) Bayesian sparse reduced rank multivariate regression, in review.

Gmez-Girldez, P.J., C. Aguilar, M. J. Polo. 2014. Natural vegetation covers as indicators of the soil water content in a semiarid mountainous watershed. Ecological Indicators 46: 524?535.

Guisan A, and Thuiller W (2005) Predicting species distribution: offering more than simple habitat models. Ecology Letters 8, 993?1009.

Guisan A. and C. Rahbek 2011. SESAM: a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. Journal of Biogeography 38, 1433?1444.

Harris, D.J., (2015) Generating realistic assemblages with a joint species distribution model, Methods in Ecology and Evolution, early on-line.

Haslett J, Whiley M, Bhattacharya S. 2006. Bayesian palaeoclimate reconstruction. Journal of the Royal Statistical Society, Series A 169:395-438.

Hooten, M.B. and N. T. Hobbs (2015) A guide to Bayesian model selection for ecologists. Ecological Monographs, 85, 3?28.

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, In press.

Iverson, L. R. and A. M. Prasad. (1998) Predicting abundance of 80 tree species following climate change in the eastern United States. Ecological Monographs, 68, 465-485.

Jackson, B.K. and Sullivan, S.M.P. 2009. Influence of wildfire severity on riparian plant community heterogeneity in an Idaho, USA wilderness. Forest Ecology and Management 259: 24?32.

Lambert, D. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34, 1-14.

Larson, M.A.; Thompson, F.R., III; Millspaugh, J.J.; Dijak, W.D.; Shifley, S.R. 2004. Linking population viability, habitat suitability, and landscape simulation models for conservation planning. Ecological Modeling. 180: 103?118.

Latimer, A.M., S. Banerjee, H. Sang, E. Mosher and J.A. Silander (2009) Hierarchical models facilitate spatial analysis of large data sets: A case study on invasive plant species in the northeastern United States. Ecology Letters, 12, 144-154.

Lauber, C.L., M. Hamady, R. Knight and N. Fierer. (2009) Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. Applied Environmental Microbiology. 75, 5111-5120.

Lavorel, S., K. Grigulis, S. McIntyre, N. S. G. Williams, D. Garden, J. Dorrough, S. Berman, F. Qutier, A. Thbault, and A. Bonis. (2008) Assessing functional diversity in the field?methodology matters! Functional Ecology, 22, 134?147.

Lawrence, E., D. Bingham, C. Liu and V. N. Nair (2008) Bayesian inference for multivariate ordinal data using parameter expansion, Technometrics, 50, 182-191.

Leininger TJ, Gelfand AE, Allen JM, and Silander JA (2013) Spatial regression modeling for compositional data with many zeros, Journal of Agricultural, Biological, and Environmental Statistics, 18, 314-334.

Li, H. 2015. Microbiome, metagenomics, and high-dimensional compositional data analysis. Annual Review Statistical Applications, 2:73-94.

Mandal, S., Van Treuren, W., White, R. A., Eggesb, M., Knight, R., and Peddada, S. D. 2015. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microbial Ecology in Health and Disease, 26, 10.3402/mehd.v26.27663. http://doi.org/10.3402/mehd.v26.27663

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J. and Possingham, H. P. (2005), Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecology Letters, 8: 1235-1246.

McCullagh, P. (1980), Regression Models for Ordinal Data, Journal of the Royal Statistical Society, Ser. B, 42, 109?142.

McMurdle, P.J. and S. Holmes. 2014. Waste not, want not: why rarifying microbiome data is inadmissible. Plos Computational Biology DOI: 10.1371/journal.pcbi.1003531.

Mokany K, Ferrier S. 2011. Predicting impacts of climate change on biodiversity: a role for semi-mechanistic community-level modelling. Diversity and Distributions 17: 374-380.

Mokany K, Harwood TD, Overton JM, Barker GM, Ferrier S. 2011. Combining alpha- and beta-diversity models to fill gaps in our knowledge of biodiversity. Ecology Letters 14:

1043-1051.

Mokany K, Harwood TD, Williams KJ, Ferrier S. 2012. Dynamic macroecology and the future for biodiversity. Global Change Biology 18: 3149-3159.

Mueller-Dombois, D. and H. Ellenberg (1986) Aims and Methods of Vegetation Ecology. Wiley, NY.

Nichols, J.D., Williams, B.K. 2006. Monitoring for conservation. Trends in Ecology and Evolution. 21, 668-673.

Novotny, V. and Y. Bassett. (2000) Rare species in communities of tropical insect herbivores: pondering the mystery of singletons. Oikos, 89, 564-572.

Ollinger SV, Smith ML, Martin ME, Hallett RA, Goodale CL, Aber JD. 2002. Regional variation in foliar chemistry and N cycling among forests of diverse history and composition. Ecology 83, 339-355.

Oksanen J. 2008 Multivariate analysis of ecological communities in R: vegan tutorial. `http://cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf`.

Ovaskainen, O., J. Hottola, and J. Siitonen (2010) Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. Ecology, 91, 2514-2521.

Ovaskainen, O., and J. Soininen. (2011) Making more out of sparse data: hierarchical modeling of species communities. Ecology 92, 289-295.

Paciorek CJ, McLachlan JS. 2009. Mapping ancient forests: Bayesian inference for spatio-temporal trends in forest composition using the fossil pollen proxy record. Journal of the American Statistical Association. 104:608-622.

Page, H. M., Culver, C. S., Dugan, J. E., and Mardian, B. 2008. Oceanographic gradients and patterns in invertebrate assemblages on offshore oil platforms. ICES Journal of Marine Science, 65: 851?861.

Parris, K. (2006) Urban amphibian assemblages as metacommunities. Journal of Animal Ecology, 75, 757-764.

Paulson J, Stine O, Bravo H, Pop M. 2013. Differential abundance analysis for microbial marker-gene surveys. Nat. Methods 10:1200-1212.

Pati, D., A Bhattacharya, NS Pillai, and D Dunson (2014) Posterior contraction in sparse Bayesian factor models for massive covariance matrices. Annals of Statistics 42, 1102-1130.

Podani, J. (2005), Multivariate exploratory analysis of ordinal data in ecology: Pitfalls, problems and solutions. Journal of Vegetation Science, 16: 497?510. doi: 10.1111/j.1654-

1103.2005.tb02390.x

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

Preston, F. W. (1948) The commonness, and rarity, of species. Ecology, 29, 254-283.

R Development Core Team (2012) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org.

Rainford SK, and Blossey B. (2014) Community-weighted mean functional effect traits determine larval amphibian responses to litter mixtures. Oecologia, 174, 1359-66.

Rajaratnam, B., S. Roberts, D. Sparks, and O. Dalal (2015) Lasso regression: estimation and shrinkage via limit of Gibbs sampling. Journal of the Royal Statistical Society: Series B (Statistical Methodology), (to appear)

Robinson, T., D. Rogers, and B. Williams. 1997. Mapping tsetse habitat suitability in the common fly belt of southern Africa using multivariate analysis fo climate and remotely sensed vegetation. Medical and Veterinary Entomology 11, 235-245.

Royle, J.A. 2004. N-mixture models for estimating population size from spatially replicated counts. Biometrics 60, 108-115.

Sahu, S.K., A.E. Gelfand, and D.M. Holland 2010. Fusing Point and Areal Level Space-time Data with Application to Wet Deposition, Journal of the Royal Statistical Society - C , 59,1, 77-103.

Saucedo-Garca A, Anaya AL, Espinosa-Garca FJ, Gonzlez MC (2014) Diversity and communities of foliar endophytic fungi from different agroecosystems of Coffea arabica L. in two regions of Veracruz, Mexico. PLoS One 9:1-11.

Schliep, E.M., A.E. Gelfand, J.S. Clark, K. Zhu. 2015. Modeling change in forest biomass across the eastern US. Environmental and Ecological Statistics, in press.

Silfver, T., Paaso, U., Rasehorn, M., Rousi, M., and Mikola, J. 2015. Genotype herbivore effect on leaf litter decomposition in Betula pendula saplings: Ecological and evolutionary consequences and the role of secondary metabolites. PLoS ONE, 10(1), e0116806. http://doi.org/10.1371/journal.pone.0116806.

Sinclair, S. J., M. D. White, and G. R. Newell. 2010. How useful are species distribution models for managing biodiversity under future climates? Ecology and Society 15(1): 8. http://www.ecologyandsociety.org/vol15/iss1/art8/

Taylor-Rodriguez, D., K. Kaufeld, E. Schliep, J. S. Clark, and A. Gelfand, 2016. Joint Species distribution modeling: dimension reduction using Dirichlet processes. Bayesian Analysis, in press.

ter Braak, C. J. F., and I. C. Prentice. (1988) A theory of gradient analysis. Advances in Ecological Research, 18, 271-313.

Thorson, J.T., M. D. Scheuerell, A. O. Shelton, K. E. See, H. J. Skaug and K. Kristensen. (2105) Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range. Methods in Ecology and Evolution, on-line view.

Tobin, J. 1958. Estimation of relationships for limited dependent variables. Econometrica 26: 24?36.

USDA Forest Service. (2012) Forest Inventory and Analysis: Fiscal year 2011 business report. WO-FS-999. U.S. Department of Agriculture, Forest Service. Washington, DC.

van Bodegom, P.M., J. C. Douma, and L. M. Verheijen, (2014) A fully traits-based approach to modeling global vegetation distribution, Proceedings of the National Academy of Sciences 111, 13733?13738.

Van der Marrel, E. and J. Franklin. (2013) Vegetation Ecology. Wiley.

Ver Hoef, J.M., and Peter L. Boveng (2007). Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data? Ecology, 88, 2766?2772.

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

Whittaker, R.H. 1978. Classification of Plant Communities. Handbook of Vegetation Science, Kluwer Academic Publishers, ISBN 90-6193-566-0.

Wieder, W.R., C. C. Cleveland, P. G. Taylor, D. R. Nemergut, E-L Hinckley, L. Philippot, D. Bru, S. R. Weintraub, M. Martin and A. R. Townsend (2013) Experimental removal and addition of leaf litter inputs reduces nitrate production and loss in a lowland tropical forest. Biogeochemistry, 13, 629-642.

Woudenberg SW, Conkling BL, O'Connell BM, LaPoint EB, Turner JA, et al. (2010) The Forest Inventory and Analysis Database: Database description and users manual version 4.0 for Phase 2. USDA Forest Service General Technical Report RMRS-GTR-245. Rocky Mountain Research Station, Fort Collins, CO.

Zhang, X., W.J. Boscardin, and T.R. Belin. 2008. Bayesian analysis of multivariate nominal measures using multivariate multinomial probit models. Computational Statistics and Data Analysis 52, 3697-3708.

Zhu, K, C. W. Woodall, S. Ghosh, A. E. Gelfand, and J. S. Clark. (2014) Dual impacts of climate change: forest migration and turnover through life history. Global Change Biology, 20, 251-264.

Tables

Table 1: Effort for discrete counts

$y_{is} = z_{is}$	E_i	w_{is}	k	$\mathbf{p}_{ik}$[1]
per plot[2]	plot area	per area	interval	partition
10	0.1 ha	100 ha^{-1}	10	$(95, 105]$
100	1.0 ha	100 ha^{-1}	100	$(99.5, 100.5]$
per OTU[3]	total reads	fraction	interval	partition
10	100	0.1	10	$(0.095, 0.105]$
10,000	100,000	0.1	10,000	$(0.099995, 0.100005]$

[1] From eqn 4

[2] e.g., plants counted on sample plots

[3] e.g., OTUs read in microbiome data

Table 2: Effort effect on partition for plot data

Data type	Partition $\mathcal{P}$	Censored intervals $\mathcal{C}$
Presence-absence PA	$\mathbf{p} = (-\infty, 0, \infty)$	$\{0, 1\}$
Continuous abundance CA	$\mathbf{p} = (-\infty, 0, \infty)$	$\{0\}$
Discrete abundance DA	$\mathbf{p}_i = (-\infty, \frac{1}{2E_i}, \frac{3}{2E_i}, \ldots, \frac{max_s(y_{is})-1/2}{E_i}, \infty)$	$\{0, 1, \ldots, max_s(y_{is})\}$
Ordinal counts OC	$\mathbf{p}_s = (-\infty, 0, p_{s,2}, p_{s,3}, \ldots, \infty)$ [1]	$\{0, 1, \ldots, K\}$
Categorical CAT	$\mathbf{p}_{is} = (-\infty, max_{k'}(w_{is,k'}), \infty)$ [2]	$\{0, 1\}$
Count composition CC	$\mathbf{p}_i = (-\infty, \frac{1}{2E_i}, \frac{3}{2E_i}, \ldots, 1 - \frac{1}{2E_i}, \infty)$	$\{0, 1, \ldots, E_i\}$
Fractional composition FC	$\mathbf{p}_i = (-\infty, 0, 1, \infty)$	$\{0, 2\}$

[1] $max_i(w_{is}|z_{is} = k) < p_{s,k} < min_i(w_{is}|z_{is} = k + 1)$

[2] $k' \in \{k|y_{is,k} = 0\}$, i.e., the maximum $w_{is,k}$ for the unobserved levels k

Figure legends

Figure 1. Zero dominance in three data types. a) Seedling hosts ($n = 762$) can be in 'morbid' or 'healthy' states, scored as 0 and 1. b) Composition count data for their endophytic microbiome ($S = 175$ OTUs occurred in at least 100 observations) are 96% zeros. c) Continuous abundance with point mass at zero–the biomass taken over $S = 98$ species on $n = 1617$ 1-ha aggregate plots is 82% zeros. (a and b from Hersh, Benetiz, Vilgalys, and Clark, in prep.)

Figure 2. A comparison of correlation values on the observation scale Y vs a latent variable W at the first stage of a hierarchical model with (a) log link, $Y = e^W$, and (b) multivariate logit link, as used for composition data, $Y_s = exp(W_s)/(1 + \sum_{s=1}^{S-1} exp(W_s)$. The reference species S has link $Y_S = 1/(1 + \sum_{s=1}^{S-1} exp(W_s)$. There are $S = 30$ species having multivariate normal distribution on the w scale, i.e., the scale where the covariance is modeled. Agreement with the observation scale would have points on the diagonal.

Figure 3. GJAM includes continuous W and discrete label Z for each observed Y. When the observation Y (vertical axis) is continuous it is equal to W. When the observation Y is discrete it is assigned to a discrete interval with label Z. The partition $\{p_k\}_{k=0}^{K-1}$ (labels on horizontal axis) defines each interval Z in terms of W. Miss-classification occurs when Z is wrong (e.g., zero inflation in c). The portion of the composition link (f) beyond point a is exaggerated in the figure for clarity and discussed in the Appendix S1. Partition points must be inferred when the scale is unknown, in which case they have a density. For ordinal data, $p_0 = -\infty$ and $p_1 = 0$. Additional partition points are estimated, each with a marginal posterior distribution in g.

Figure 4. Censoring in gjam. As a data-generating model (a), a realization W that lies within a censored interval is translated by the partition $\mathbf{p}$ to discrete Y. The distribution of data (bars at left) is induced by the latent scale and the partition, shown as horizontal bars. For inference (b), observed discrete Y takes values on the latent scale from a truncated distribution.

Figure 5. Mean-variance relationships. a) Interval censoring controls variance, which increases with partition width (shown as vertical dashed lines at 0, 1, 2, 4, 8, 16). Intervals are shown for the predictive mean values of $\hat{\mathbf{Y}}$ b) For composition-count (microbiome) data partition width declines with total counts for the sample, thus decreasing variance with increasing effort.

Figure 6. Joint modeling of simulated data for $Q - 1 = 4$ predictors, $n = 2000$ observations,

and $S = 17$ species. Data types include continuous with no zero censoring (CON), presence-absence (PA), continuous abundance (CA), discrete abundance (DA), count composition (CC), and ordinal counts (OC). Coefficient estimates in (a) and correlation estimates in (b) include all combinations of data types. For ordinal categories partitions are accurately predicted in (c). Vertical whiskers are 95% credible intervals.

Figure 7. Joint data prediction for the example in Figure 6. Frequency of observations in $\mathbf{Y}$ is shown at the base of graphs. Box and whisker plots are 68% and 95% predictive intervals.

Figure 8. Parameter estimates $(\mathbf{B}, \mathbf{R})$ and data prediction $(\mathbf{Y})$ for abundance data fitted as abundance (left) and as presence/absence (right). For this simulated example $n = 200$, $S = 10$, $Q = 5$. Each panel includes means and 95% intervals. Both analyses were done with the GJAM based on the same simulated abundance data. For the presence-absence example, matrix $\mathbf{B}$ is translated to the correlation scale (Appendix S1).

Figure 9. GLM and GJAM predictions for (a) host status from Figure 1a and for (b) stem counts, for the same plots represented by biomass data in Figure 1c. GLMs use a Bernoulli likelihood with a probit link and a Poisson likelihood with log link, respectively. In (a) predictor variables are temperature, host species, and polyculture treatment, the last two variables being factors. In (b) the predictors are stand age, temperature, moisture, climatic deficit, topography, and soils, the last being a factor. The 1:1 line of agreement and root mean square prediction error (rmspe) are shown for each example. Data in (a) from Hersh, Benetiz, Vilgalys, and Clark, in preparation.

Figure 10. Predicted continuous foliar traits (a), biomass (b), and species richness (c) for the FIA example. The distribution of data is shown as histograms. Boxes and whiskers are 68% and 95% predictive intervals.

Figure 11. Sensitivity $\hat{\mathbf{F}}$ from eqn (8) (left) and coefficient matrix $\hat{\mathbf{B}}$ (right) for the FIA example. The diagonal of $\hat{\mathbf{F}}$ is the sensitivity vector $\hat{\mathbf{f}}$ (eqn 9), showing large values for slope (u_1) and two soil types, resulting from strong effects of these variables in the $\mathbf{B}$ matrix at right. Predictor variables described in the Appendix S1 include temperature, moisture, four soil types (a multilevel factor), and topography, the latter including $u_1 = \sin(\mathit{slope})$, $u_2 = \sin(\mathit{slope})\sin(\mathit{aspect})$, and $u_3 = \sin(\mathit{slope})\cos(\mathit{aspect})$ (Clark 1990). The heat color scale is strong negative (blue) to zero (white) to red (strong positive).

Figure 12. Effect of aspect on basal area for species showing the greatest responses, given as the sum $\beta_{u_1,s}u_1 + \beta_{u_2,s}u_2 + \beta_{u_3,s}u_3$. Envelopes bound responses for slopes of $10 - 20°$. The vertical scale is in units of basal area (m^2 ha^{-1}).

Figure 13. Inverse prediction of a) temperature, b) moisture, c) slope, and c) aspect. In d symbol size is proportional to slope (zero slope has no aspect). Boxes and whiskers are 68% and 95% predictive intervals. The distribution of data is shown as historgrams.

Figure 14. Correlation structure in data (a) and in response to the environment (b). The structure in (a) comes from the ordering of species by cluster analysis of the abundance data. Predictive distributions for the matrix $\hat{\mathbf{E}}$ in (b) are ordered as in (a), but show no such structure. When clustered instead by $\hat{\mathbf{E}}$ clear structure emerges (c).

Figure 15. Reads per OTU massively overrepresents zeros, but can range as high as 10^6.

Figure 16. Inverse prediction of $\mathbf{X}$ from soil microbiome data show poor prediction for sample latitude (a) and pH (b), but good prediction of many habitats (c), a multilevel factor in the model. The 'reference' category refers to habitats that were rare in the data. Missing covariate values are shown as blue dots at right of (a) and (b). the relative number of samples in each habitat category are shown with shading at the base of (c).

Figure 17. Response matrix $\hat{\mathbf{E}}$ showing groups of OTUs similar in their responses to environmental variables, explained primarily by the factor habitat in the coefficient matrix $\mathbf{B}$ (names in green at right).

Figures

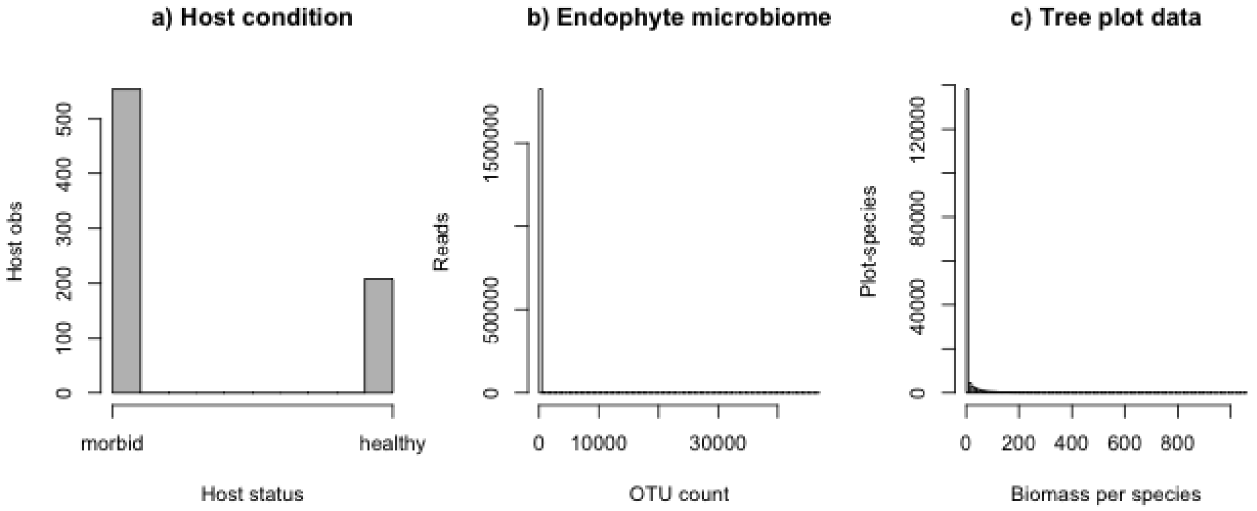

Figure 1: Zero dominance in three data types. a) Seedling hosts ($n = 762$) can be in 'morbid' or 'healthy' states, scored as 0 and 1. b) Composition count data for their endophytic microbiome ($S = 175$ OTUs occurred in at least 100 observations) are 96% zeros. c) Continuous abundance with point mass at zero–the biomass taken over $S = 98$ species on $n = 1617$ 1-ha aggregate plots is 82% zeros. (a and b from Hersh, Benetiz, Vilgalys, and Clark, in prep.)

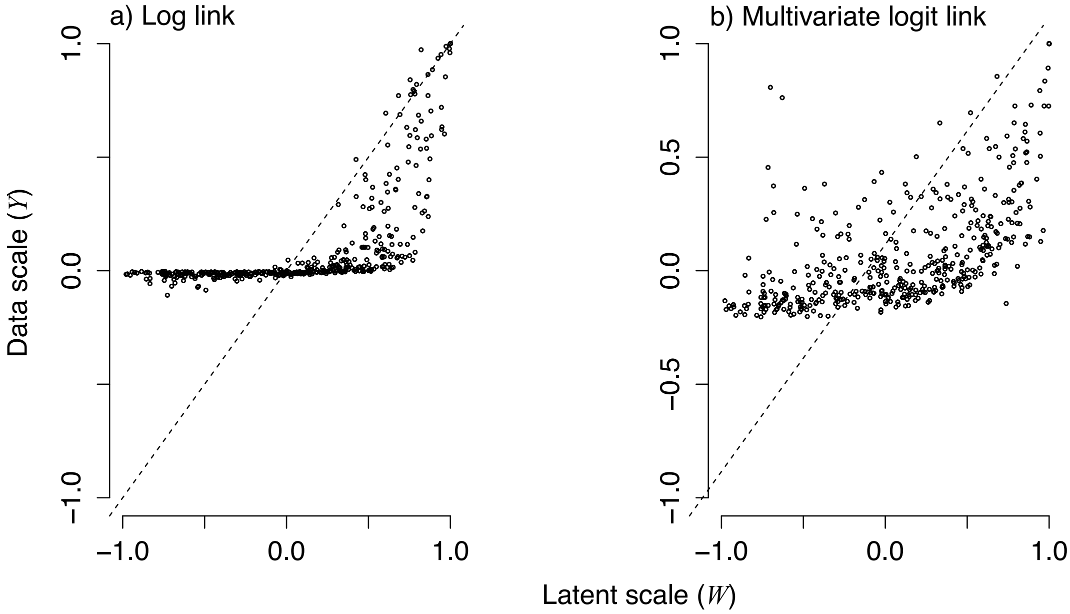

Figure 2: A comparison of correlation values on the observation scale Y vs a latent variable W at the first stage of a hierarchical model with (a) log link, $Y = e^W$, and (b) multivariate logit link, as used for composition data, $Y_s = exp(W_s)/(1 + \sum_{s=1}^{S-1} exp(W_s)$. The reference species S has link $Y_S = 1/(1 + \sum_{s=1}^{S-1} exp(W_s)$. There are $S = 30$ species having multivariate normal distribution on the w scale, i.e., the scale where the covariance is modeled. Agreement with the observation scale would have points on the diagonal.

.

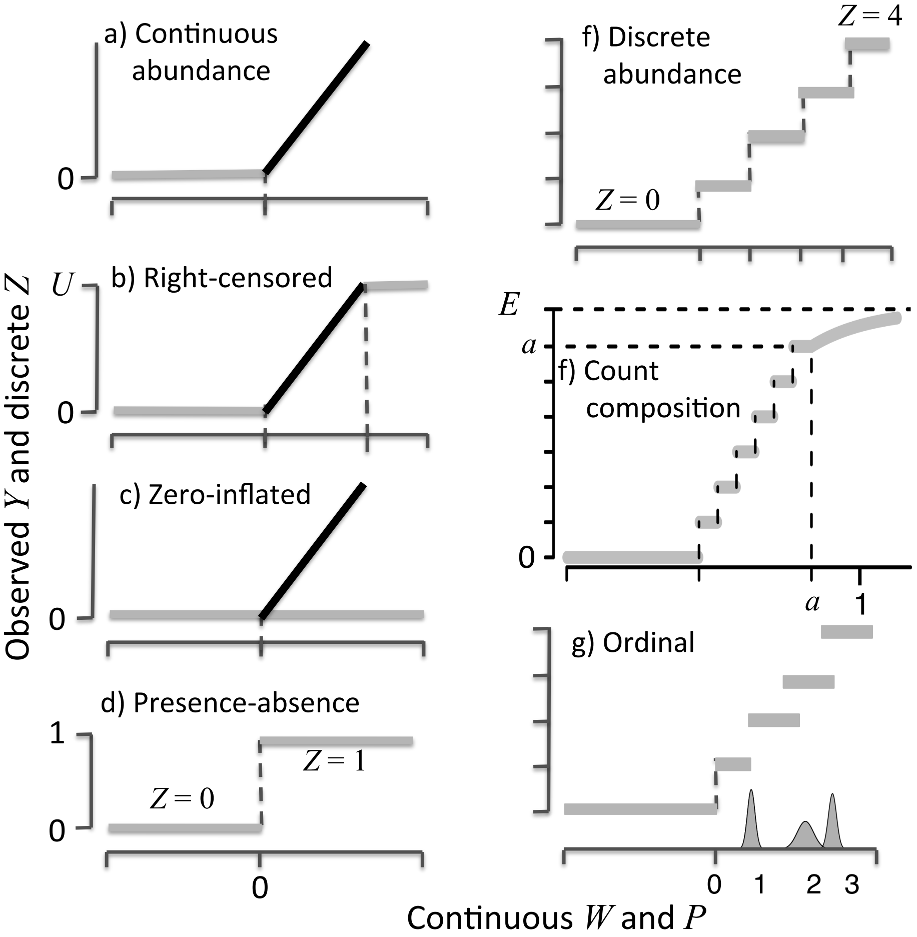

Figure 3: GJAM includes continuous W and discrete label Z for each observed Y. When the observation Y (vertical axis) is continuous it is equal to W. When the observation Y is discrete it is assigned to a discrete interval with label Z. The partition $\{p_k\}_{k=0}^{K-1}$ (labels on horizontal axis) defines each interval Z in terms of W. Miss-classification occurs when Z is wrong (e.g., zero inflation in c). The portion of the composition link (f) beyond point a is exaggerated in the figure for clarity and discussed in the Appendix S1. Partition points must be inferred when the scale is unknown, in which case they have a density. For ordinal data, $p_0 = -\infty$ and $p_1 = 0$. Additional partition points are estimated, each with a marginal posterior distribution in g.

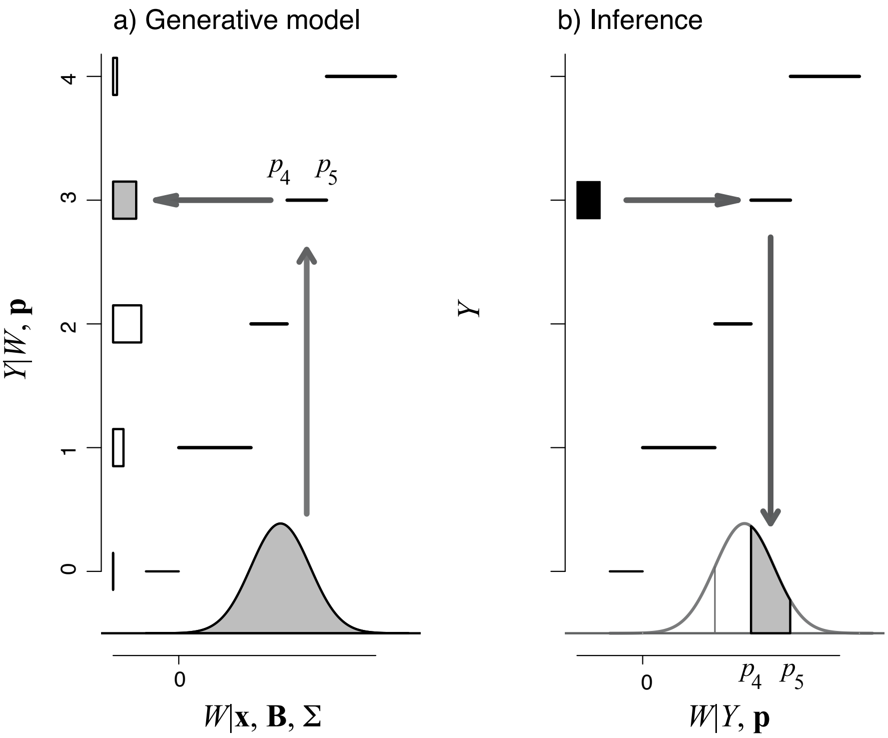

Figure 4: Censoring in gjam. As a data-generating model (a), a realization W that lies within a censored interval is translated by the partition $\mathbf{p}$ to discrete Y. The distribution of data (bars at left) is induced by the latent scale and the partition, shown as horizontal bars. For inference (b), observed discrete Y takes values on the latent scale from a truncated distribution.

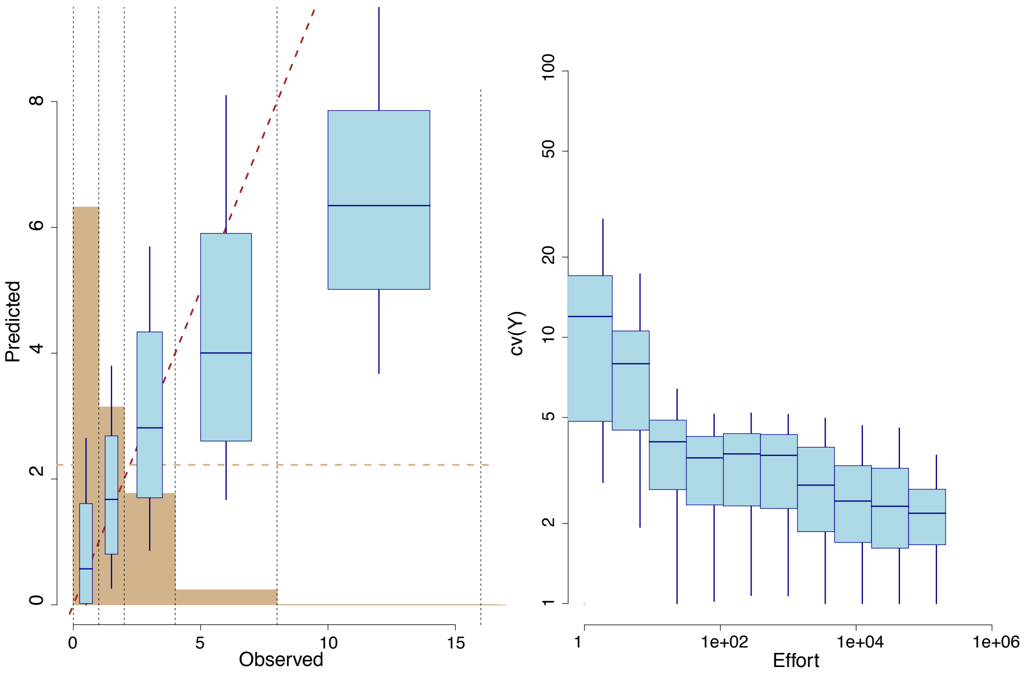

Figure 5: Mean-variance relationships. a) Interval censoring controls variance, which increases with partition width (shown as vertical dashed lines at 0, 1, 2, 4, 8, 16). Intervals are shown for the predictive mean values of $\hat{\mathbf{Y}}$ b) For composition-count (microbiome) data partition width declines with total counts for the sample, thus decreasing variance with increasing effort.

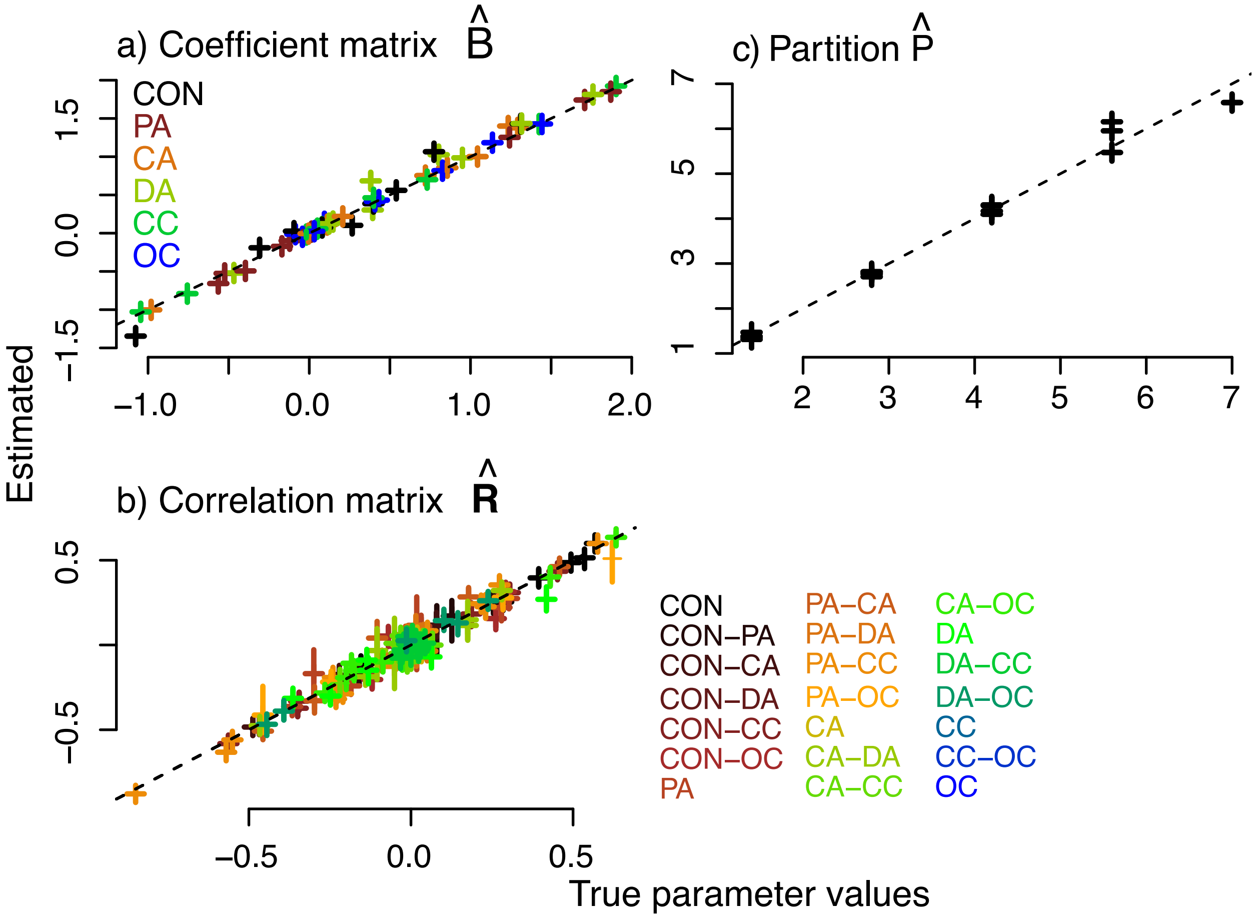

Figure 6: Joint modeling of simulated data for $Q-1=4$ predictors, $n=2000$ observations, and $S=17$ species. Data types include continuous with no zero censoring (CON), presence-absence (PA), continuous abundance (CA), discrete abundance (DA), count composition (CC), and ordinal counts (OC). Coefficient estimates in (a) and correlation estimates in (b) include all combinations of data types. For ordinal categories partitions are accurately predicted in (c). Vertical whiskers are 95% credible intervals.

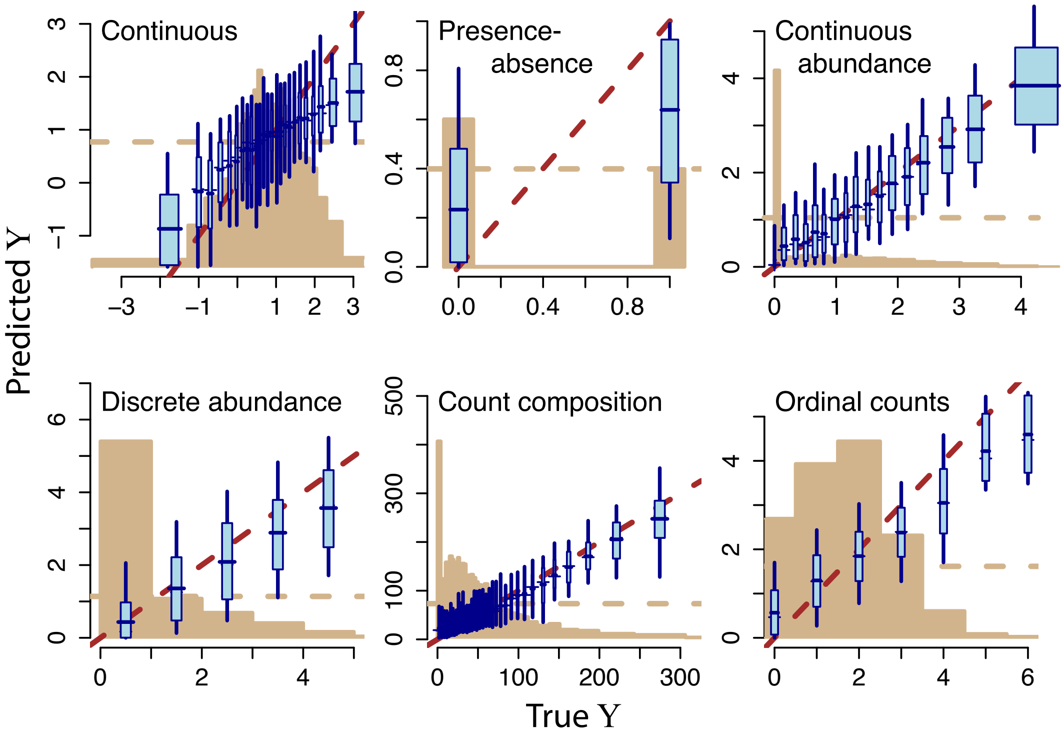

Figure 7: Joint data prediction for the example in Figure 6. Frequency of observations in **Y** is shown at the base of graphs. Box and whisker plots are 68% and 95% predictive intervals.

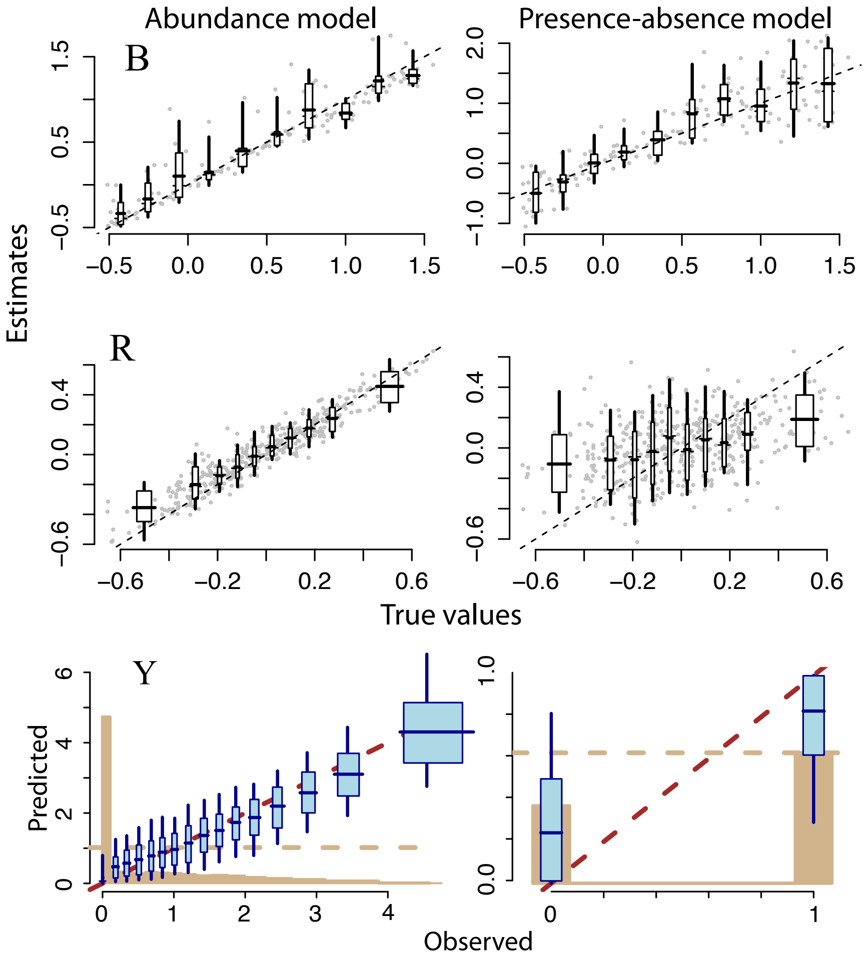

Figure 8: Parameter estimates $(\mathbf{B}, \mathbf{R})$ and data prediction $(\mathbf{Y})$ for abundance data fitted as abundance (left) and as presence/absence (right). For this simulated example $n = 200$, $S = 10$, $Q = 5$. Each panel includes means and 95% intervals. Both analyses were done with the GJAM based on the same simulated abundance data. For the presence-absence example, matrix $\mathbf{B}$ is translated to the correlation scale (Appendix S1).

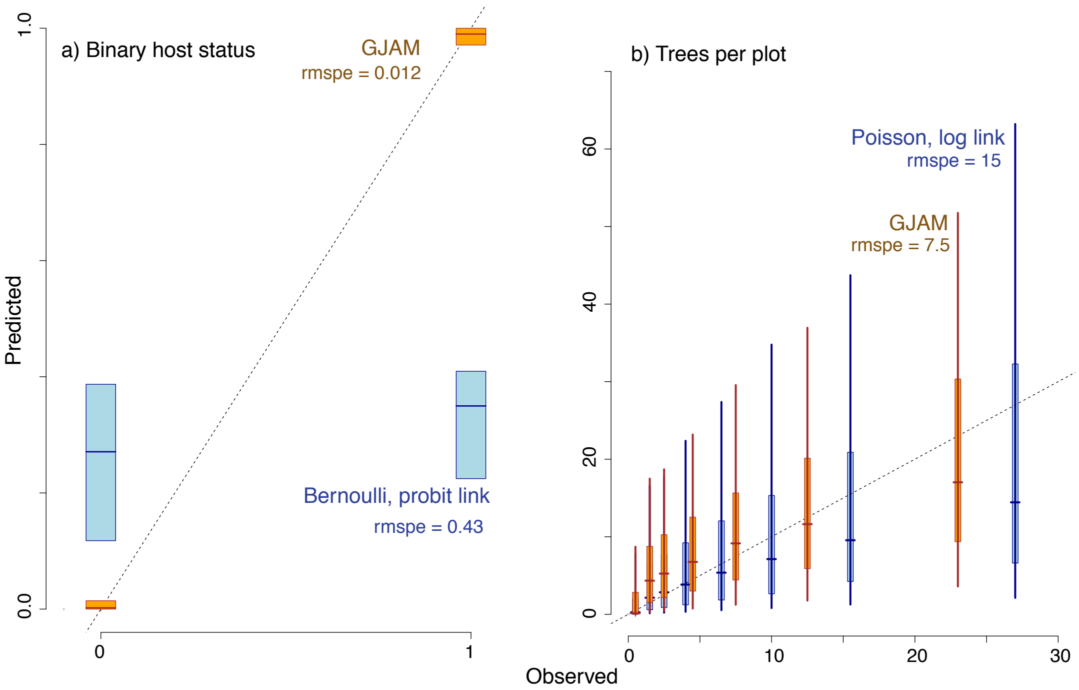

Figure 9: GLM and GJAM predictions for (a) host status from Figure 1*a* and for (b) stem counts, for the same plots represented by biomass data in Figure 1*c*. GLMs use a Bernoulli likelihood with a probit link and a Poisson likelihood with log link, respectively. In (a) predictor variables are temperature, host species, and polyculture treatment, the last two variables being factors. GJAM models the combined host status and microbiome as responses. In (b) the predictors are stand age, temperature, moisture, climatic deficit, topography, and soils, the last being a factor. The 1:1 line of agreement and root mean square prediction error (rmspe) are shown for each example. Data in (a) from Hersh, Benetiz, Vilgalys, and Clark, in preparation.

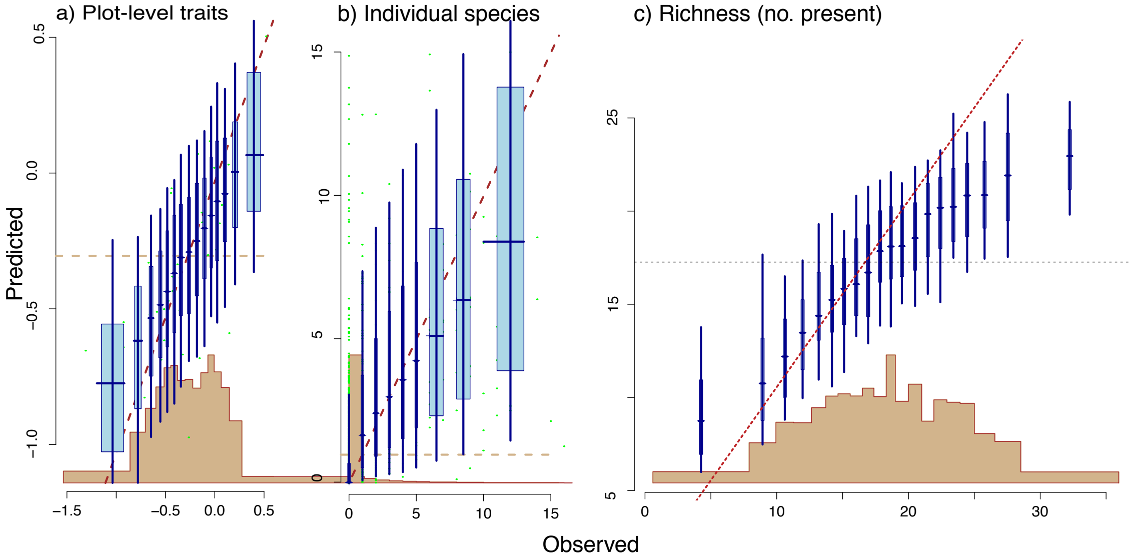

Figure 10: Predicted continuous foliar traits (both N and P) (a), biomass (b), and species richness (c) for the FIA example. The distribution of data is shown as histograms. Boxes and whiskers are 68% and 95% predictive intervals.

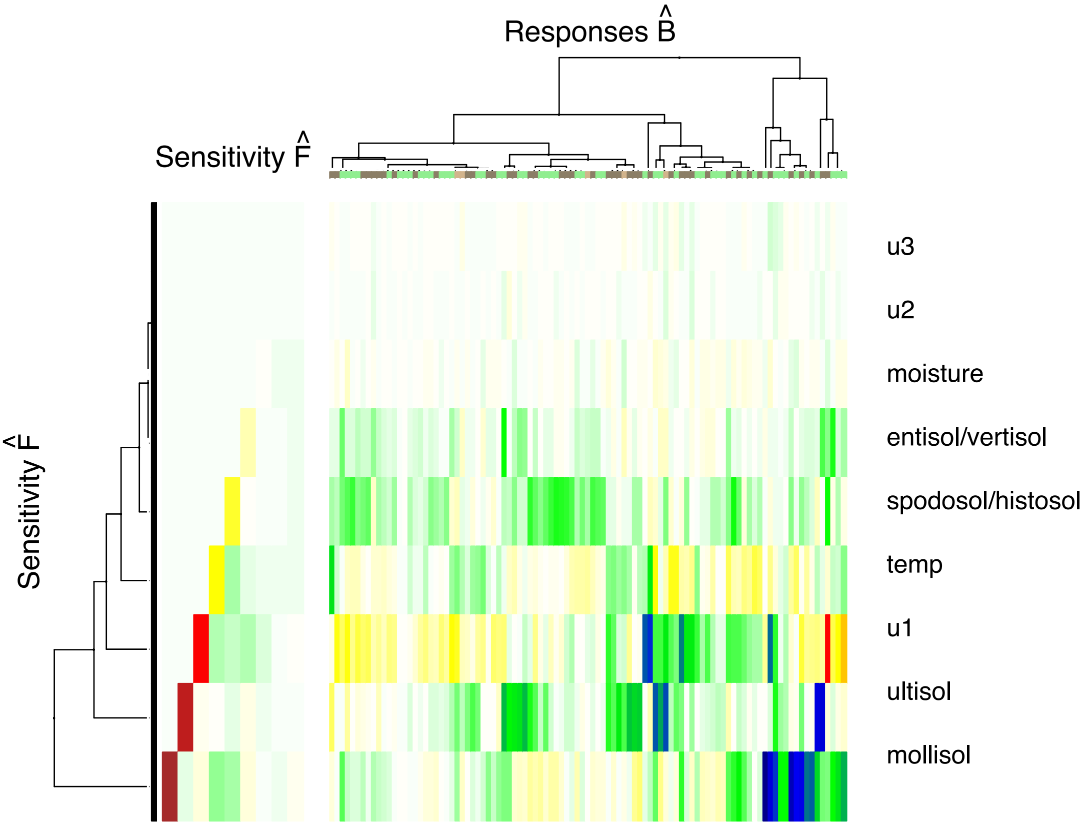

Figure 11: Sensitivity $\hat{\mathbf{F}}$ from eqn (8) (left) and coefficient matrix $\hat{\mathbf{B}}$ (right) for the FIA example. The diagonal of $\hat{\mathbf{F}}$ is the sensitivity vector $\hat{\mathbf{f}}$ (eqn 9), showing large values for slope (u_1) and two soil types, resulting from strong effects of these variables in the $\mathbf{B}$ matrix at right. Predictor variables described in the Appendix S1 include temperature, moisture, four soil types (a multilevel factor), and topography, the latter including $u_1 = \sin(slope)$, $u_2 = \sin(slope)\sin(aspect)$, and $u_3 = \sin(slope)\cos(aspect)$ (Clark 1990). The heat color scale is strong negative (blue) to zero (white) to red (strong positive).

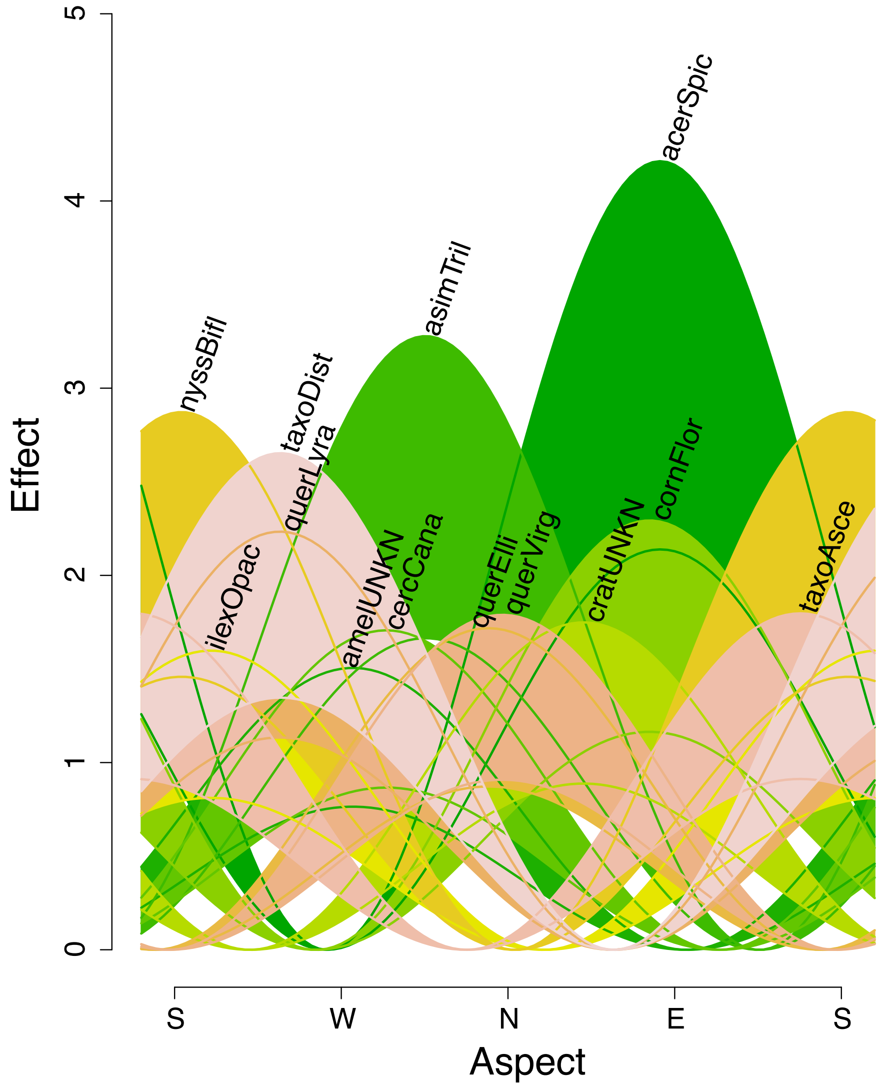

Figure 12: Effect of aspect on basal area for species showing the greatest responses, given as the sum $\beta_{u_1,s}u_1 + \beta_{u_2,s}u_2 + \beta_{u_3,s}u_3$. Envelopes bound responses for slopes of $10 - 20°$. The vertical scale is in units of basal area (m^2 ha^{-1}).

.

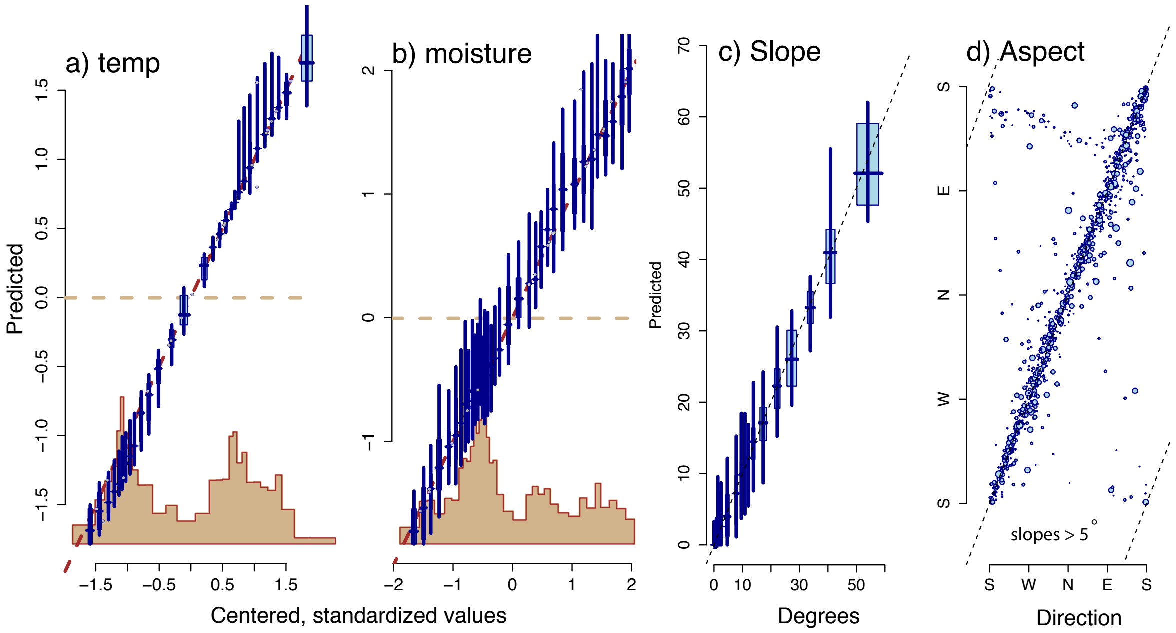

Figure 13: Inverse prediction of a) temperature, b) moisture, c) slope, and c) aspect. In d symbol size is proportional to slope (zero slope has no aspect). Boxes and whiskers are 68% and 95% predictive intervals. The distribution of data is shown as historgrams.

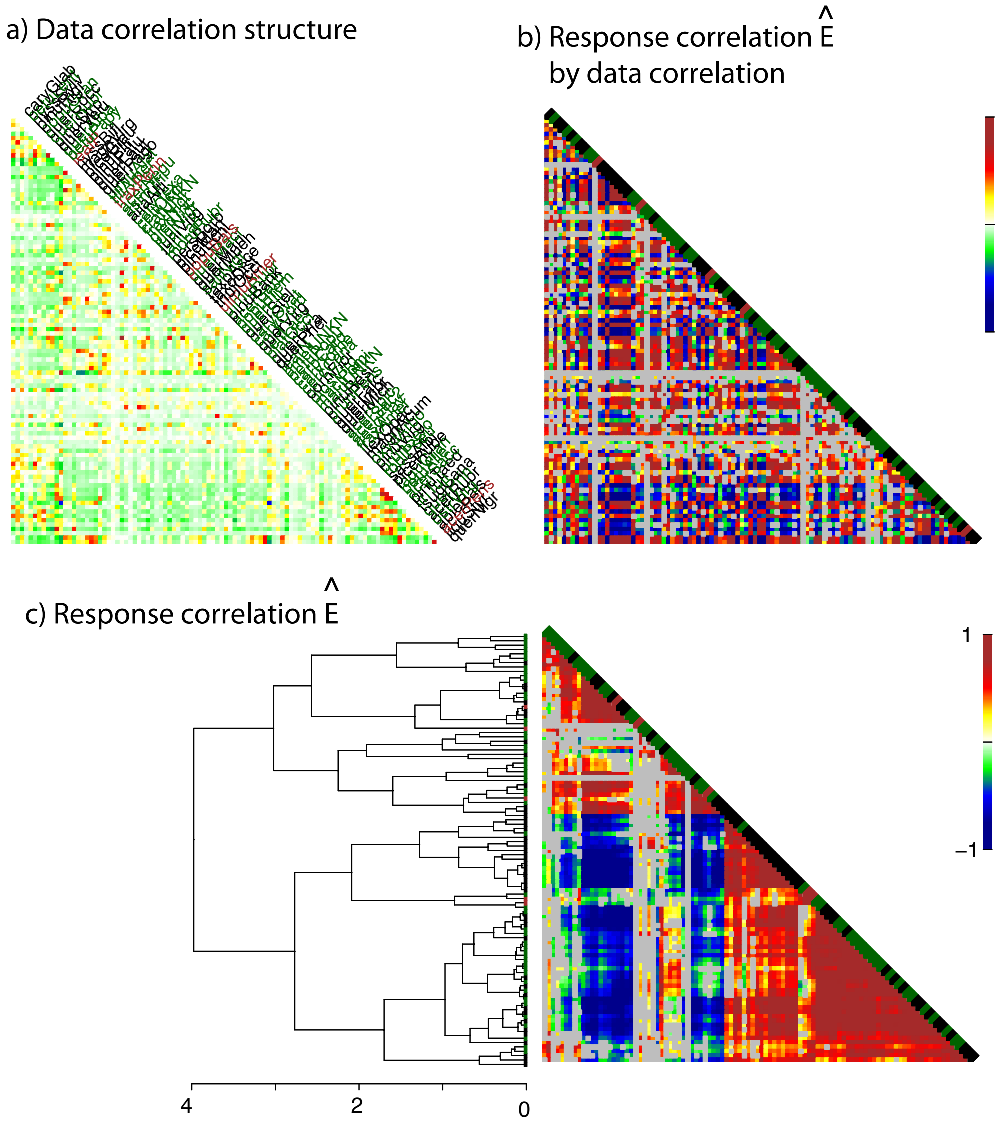

Figure 14: Correlation structure in data (a) and in response to the environment (b). The structure in (a) comes from the ordering of species by cluster analysis of the abundance data. Predictive distributions for the matrix $\hat{\mathbf{E}}$ in (b) are ordered as in (a), but show no such structure. When clustered instead by $\hat{\mathbf{E}}$ clear structure emerges (c).

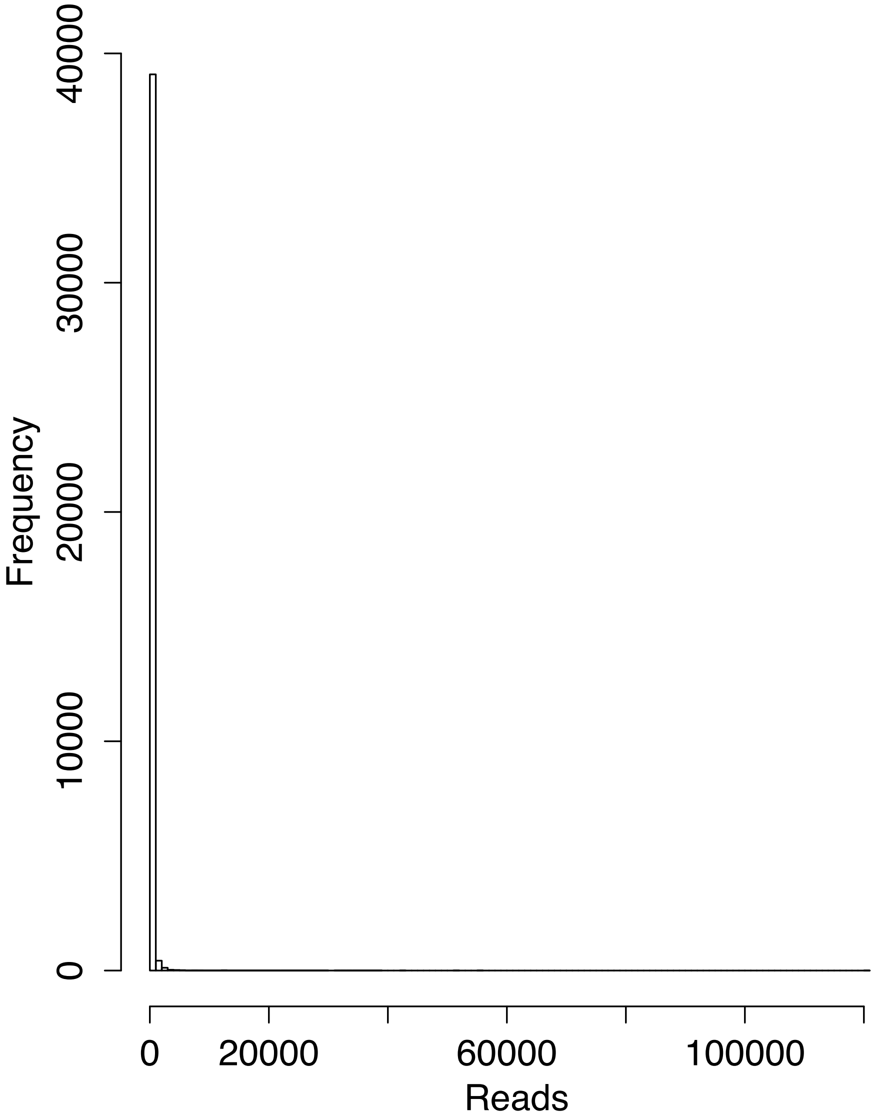

Figure 15: Reads per OTU massively overrepresents zeros, but can range as high as 10^6.

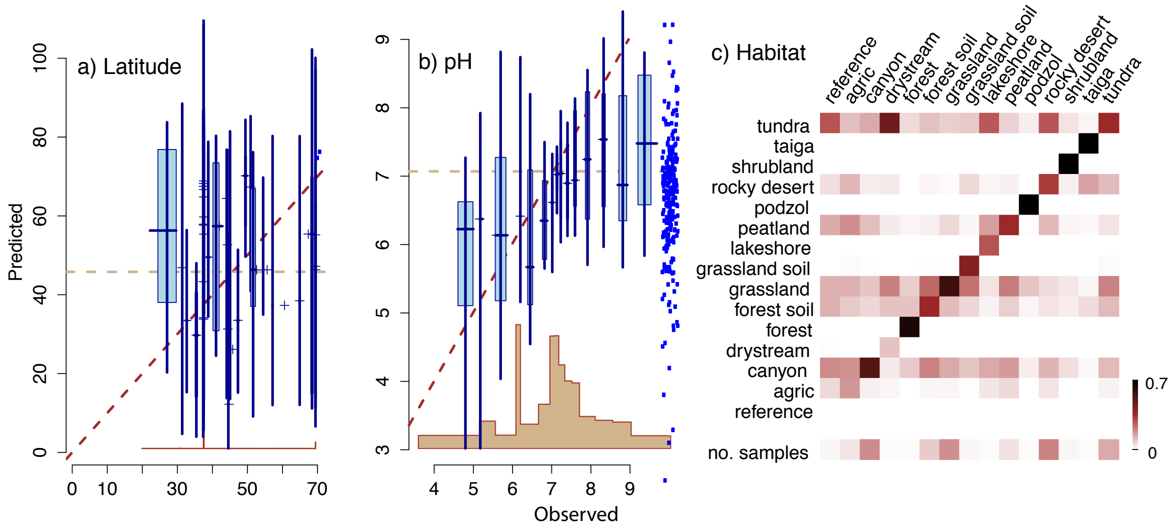

Figure 16: Inverse prediction of **X** from soil microbiome data show poor prediction for sample latitude (a) and pH (b), but good prediction of many habitats (c), a multilevel factor in the model. The 'reference' category refers to habitats that were rare in the data. Missing covariate values are shown as blue dots at right of (a) and (b). the relative number of samples in each habitat category are shown with shading at the base of (c).

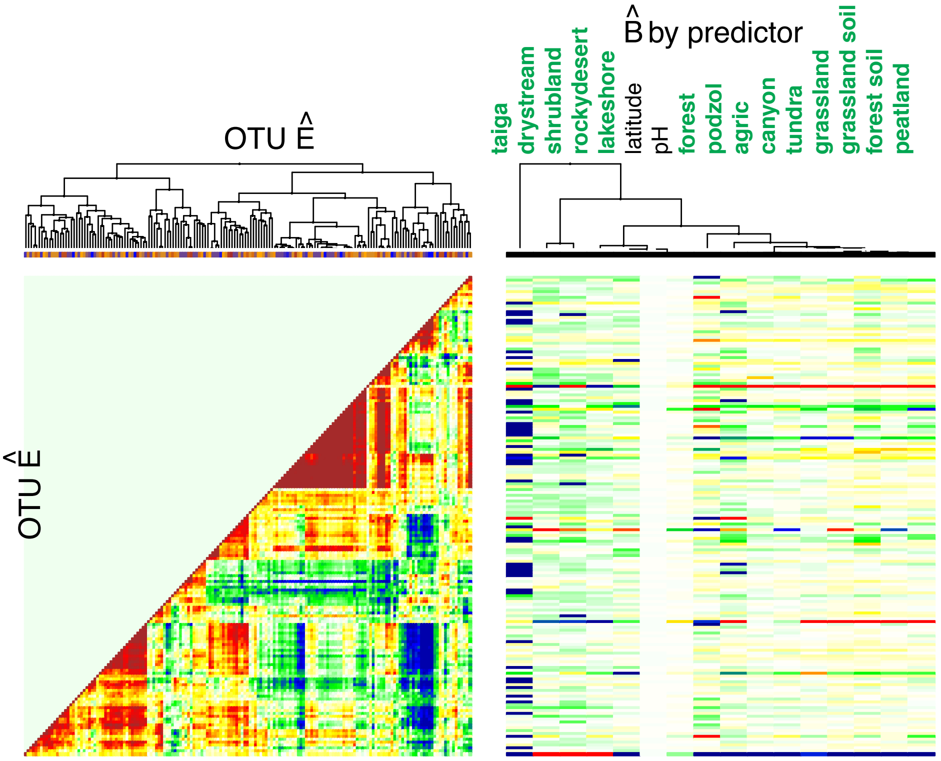

Figure 17: Response matrix $\hat{\mathbf{E}}$ showing groups of OTUs similar in their responses to environmental variables, explained primarily by the factor habitat in the coefficient matrix $\mathbf{B}$ (names in green at right).