Biostatistics (2022) **23**, 1, *pp.* 207–222
doi: 10.1093/biostatistics/kxaa015
Advance Access publication on May 20, 2020

Estimating diversity in networked ecological communities

AMY D. WILLIS*, BRYAN D. MARTIN

Department of Biostatistics and Department of Statistics, University of Washington, Health Sciences Building, 1959 NE Pacific St, Seattle WA 98195, USA

* adwillis@uw.edu

Summary

Comparing ecological communities across environmental gradients can be challenging, especially when the number of different taxonomic groups in the communities is large. In this setting, community-level summaries called *diversity indices* are widely used to detect changes in the community ecology. However, estimation of diversity indices has received relatively little attention from the statistical community. The most common estimates of diversity are the maximum likelihood estimates of the parameters of a multinomial model, even though the multinomial model implies strict assumptions about the sampling mechanism. In particular, the multinomial model prohibits ecological networks, where taxa positively and negatively co-occur. In this article, we leverage models from the compositional data literature that explicitly account for co-occurrence networks and use them to estimate diversity. Instead of proposing new diversity indices, we estimate popular diversity indices under these models. While the methodology is general, we illustrate the approach for the estimation of the Shannon, Simpson, Bray–Curtis, and Euclidean diversity indices. We contrast our method to multinomial, low-rank, and nonparametric methods for estimating diversity indices. Under simulation, we find that the greatest gains of the method are in strongly networked communities with many taxa. Therefore, to illustrate the method, we analyze the microbiome of seafloor basalts based on a 16S amplicon sequencing dataset with 1425 taxa and 12 communities.

Keywords: Diversity; Ecology; High throughput sequencing; Microbiome; Network.

1. Introduction

Microbial communities are composed of enormous numbers of different microbes, ranging from highly abundant taxa to rare taxa that are often unobserved. Data obtained from microbiome surveys often take the form of high-dimensional count data, generally with additional covariate information regarding the experimental conditions under which the communities were observed (Li, 2015). Detecting patterns in this data is challenging, partly because of its dimension. Analysis of *diversity* is a standard approach to summarizing and comparing high-dimensional community composition data in ecological studies and is ubiquitous in the microbiome literature (Callahan *and others*, 2016).

Consider a microbial community of C taxonomic groups (taxa), which are present in relative abundances $z = (z_1, \ldots, z_C)$. Depending on the ecosystem under study, C may be on the order of hundreds, but may also be in the tens of thousands or greater. An α-diversity index $f : \mathbb{S}^{C-1} \to \mathbb{R}$ summarizes z, where $\mathbb{S}^d$ is

*To whom correspondence should be addressed.

the d-dimensional simplex. Similarly, β-diversity indices $g : \mathbb{S}^{C-1} \times \mathbb{S}^{C-1} \to \mathbb{R}$ summarize information from two communities. β-diversity indices summarize between-community structure, while α-diversity indices summarize within-community structure. Specific examples are given in Section 2.

Despite the prevalence of α- and β-diversity analyses in ecology, statistical methodology to estimate these functions is relatively underdeveloped. In particular, much of the existing literature focuses on estimating diversity under the assumption of observations drawn from a multinomial distribution with unknown probability vector z (Miller, 1955; Zahl, 1977; Zhang and Zhou, 2010; Hsieh *and others*, 2016; Cao *and others*, 2019b). Fortunately, there exist sophisticated models for community composition data that permit a more flexible co-occurrence structure than that implied by the multinomial distribution. In this article, we utilize models from the compositional data literature that explicitly permit co-occurrence of taxa. The novelty of this article is in leveraging these models to estimate diversity indices, developing parametric and nonparametric variance estimates, and developing software implementing the method. Note that we do not propose novel diversity indices, but develop novel estimators of widely analyzed diversity indices.

In addition to incorporating network structure, the proposed method has a number of advantages over existing methods for diversity estimation. Most notably, while it is common in practice to estimate the diversity of each community individually, our method can pool information across multiple samples to estimate the diversity of the ecological communities from which the samples were drawn. Our methodology also permits a principled method for predicting diversity in communities that were not sampled. Our method achieves substantial improvements in estimation performance under simulation and is computationally feasible for modern microbiome datasets. The method is available as an `R` package at github.com/adw96/DivNet.

The manuscript is laid out as follows: Section 2 introduces methods for estimating α- and β-diversity. In Section 3, we introduce our model for estimating diversity, and in Section 4, we discuss estimation of the model parameters and variance. Section 5 introduces a simulation study to evaluate the performance of the method (see also supplementary material available at *Biostatistics* online), and an example of the method is discussed in Section 6. We conclude with a discussion of the method, its limitations, and avenues for future research in Section 7.

2. LITERATURE REVIEW: ESTIMATING α- AND β-DIVERSITY

Suppose that we have samples from $i = 1, \ldots, n$ communities. Let $\mathcal{C}_i$ denote the set of all taxa in community i, and let $C_i = |\mathcal{C}_i|$ denote the number of taxa in the ith community. Let $\mathcal{C} = \cup_i \mathcal{C}_i$, $Q = |\mathcal{C}|$ and $q = 1, \ldots, Q$ index the taxa. Let $Z_{iq} \in [0, 1]$ denote the (unknown) relative abundance of taxon q in community i, noting that $\sum_{q=1}^{Q} Z_{iq} = 1$. (Note that while Z_{iq} is an unknown parameter, in our model below we will treat it as a latent random variable, and so we use this notation throughout for consistency.) Associated with each community is a known vector of covariates $X_i \in \mathbb{R}^p$ where $p \geq 1$.

Suppose that from the ith community, M_i individuals are observed and classified into the Q taxonomic groups. Let W_{iq} denote the number of times that taxon q was observed in the sample from community i. Therefore, to estimate summary statistics associated with the communities, the information available on which to base estimation is $W \in \mathbb{R}^{n \times Q}$ and $X \in \mathbb{R}^{n \times p}$.

While members of an ecological community may differ in their levels of relatedness, to constrain the scope of this article we do not consider measures of diversity that are functions of taxonomy, such as Faith's phylogenetic diversity (Faith, 1992), branch weighted phylogenetic diversity (McCoy and Matsen, 2013) or UniFrac (Lozupone and Knight, 2005).

2.1. α-*Diversity*

There are a number of different α-diversity indices that are widely used in the literature. This is because different indices reflect different features of communities. Two of the most common indices are the Shannon

entropy (also called the Shannon index), and the Simpson index. While the diversity estimation framework that we will introduce is applicable to any α-diversity index that is a function of taxon abundance (including any Hill number; Hill, 1973), we will focus on the Shannon and Simpson indices to illustrate our method.

2.1.1. *Shannon entropy* One of the most common α-diversity indices is the Shannon entropy (Shannon, 1948). The Shannon index of community i is defined as

$$\alpha_{i,Shannon} = -\sum_{q \in \mathcal{C}_i} Z_{iq} \log(Z_{iq}). \tag{2.1}$$

This index captures information about both the species richness (number of species) and the relative abundances of the species: as the number of species in the population increases, so does the Shannon index, and as the relative abundances diverge from a uniform distribution and become more unequal, the Shannon index decreases (for fixed $|\mathcal{C}_i|$, the entropy is maximized when $Z_{iq} = 1/C_i$ for all $q \in \mathcal{C}_i$).

Under the model $\mathbf{W}_i \sim \text{Multinomial}(M_i, \mathbf{Z}_i)$, the maximum likelihood estimate (MLE) of $\alpha_{i,\text{Shannon}}$ is $-\sum_{q \in \mathcal{C}_i} \frac{W_{iq}}{M_i} \log\left(\frac{W_{iq}}{M_i}\right)$, with the convention that if $W_{iq} = 0$, then $\frac{W_{iq}}{M_i} \log\left(\frac{W_{iq}}{M_i}\right) \equiv 0$, since $\lim_{x \to 0} x \log x = 0$. This estimate is almost ubiquitous in the ecological literature (Weiss *and others*, 2017). The multinomial MLE of Shannon diversity is often referred to as the plug-in estimate (Vu *and others*, 2007). The multinomial MLE is negatively biased by $\frac{|C_i|-1}{2M_i} + O(M_i^2)$ (Basharin, 1959), for which various corrections have been proposed, including adding $\frac{|C_i|-1}{2M_i}$ (the *Miller-Maddow* MLE correction; Miller, 1955), and jackknifing (Zahl, 1977).

Noting that unobserved (latent) taxa are often a substantial source of error in estimating the Shannon index, Chao and Shen (2003) proposed using the Good–Turing estimate of species richness and adjusting for the missing taxa, obtaining the estimate $-\sum_{q \in \mathcal{C}_i} \frac{\hat{C}_i \hat{\pi}_{iq} \log(\hat{C}_i \hat{\pi}_{iq})}{1-(1-\hat{C}_i \hat{\pi}_{iq})^n}$, where $\hat{\pi}_{iq} = W_{iq}/M_i$ and $\hat{C}_i = 1 - \sum_q 1_{\{W_{iq}=1\}} / \sum_q W_{iq}$. Vu *and others* (2007) show that this estimator is consistent and converges with the optimal rate $O_P(1/\log(M_i))$.

More recently, Chao *and others* (2013) proposed to correct bias due to latent taxa by subsampling taxa and extrapolating from the sequentially smaller subsamples. The idea behind this method is to sample $m_1, m_2, \ldots, m_k$, microbes without replacement from the M_i observed microbes. k multinomial MLEs of the Shannon diversity are constructed based on each of the subsamples, and we call the jth estimate $\hat{\alpha}_i(m_j)$. The curve $\{m_j, \hat{\alpha}_i(m_j)\}_{j=1}^k$ is then constructed, along with an estimate of the slope of the curve. This curve is then extrapolated based on the estimated slopes to $m \to \infty$. The method is implemented in the R package `iNEXT` (Hsieh *and others*, 2016), with which we compare our method. We note that the taxa are subsampled independently to reflect the assumptions of the multinomial model. An alternative approach to adjusting for latent taxa originates in the compositional data analysis literature. To estimate the compositions Z_{iq}, Martín-Fernández *and others* (2003) propose replacing observed values of W_{ij} that are exactly zero with 0.5, and so Cao *and others* (2019b) consider the resulting *zero-replace* α-diversity estimator $-\sum_{q \in \mathcal{C}} \frac{W_{iq} \vee 0.5}{\sum_{r \in \mathcal{C}} W_{ir} \vee 0.5} \log\left(\frac{W_{iq} \vee 0.5}{\sum_{r \in \mathcal{C}} W_{ir} \vee 0.5}\right)$. Cao *and others* (2019b) also extend this idea by fitting a Poisson-Multinomial model to W via a regularization approach that penalizes the nuclear norm of Z, thereby obtaining a low-rank estimate of Z that is close to the MLE under a Poisson-Multinomial model. No publicly available software implements the low-rank matrix method.

2.1.2. *Simpson index* Simpson (1949) defined the index now known as the *Simpson index*:

$$\alpha_{i,Si} = \sum_{q \in \mathcal{C}_i} Z_{iq}^2. \tag{2.2}$$

Similar to the Shannon index, the most common estimate of the Simpson index is the plug-in estimate $\hat{\alpha}_{i,Si,\text{plug-in}} = \sum_{q \in \mathcal{C}_i} \left(\frac{W_{iq}}{M_i}\right)^2$. Zhang and Zhou (2010) demonstrated that under independent sampling from a multinomial distribution, $\frac{M_i}{M_i - 1}\hat{\alpha}_{i,Si,\text{plug-in}}$ is unbiased and asymptotically normally distributed. However, since M_i generally exceeds 1000 in microbiome studies, the difference between the Zhang and Zhou (2010) and the plug-in estimate is negligible in our setting.

A number of approaches to estimating the Shannon index are also applicable to estimating the Simpson index. For example, Cao *and others* (2019b) investigate the performance of the zero-replace and low-rank approach to estimating the Simpson index. The extrapolation approach of Hsieh *and others* (2016) also applies to the Simpson index.

2.1.3. *α-Diversity with covariates* With the exception of Cao *and others* (2019b), all of the estimates for α_i discussed above are only functions of the abundance vectors $\mathbf{W}_i$. Notably, none utilize the full abundance matrix W nor the covariate matrix X. To address this, De'ath (2012) proposed a multinomial logistic regression approach to estimating the Shannon diversity. Advantages of this method include that diversity can be extrapolated, while disadvantages include a lack of publicly available software and no generative model for the species counts. More recently, Arbel *and others* (2016) proposed a nonparametric Bayesian model that exploits structure in W as well as incorporating covariate information. Specifically, the model for the taxon counts W given the taxon relative abundances Z is nonparametric. The marginal prior distribution for Z_i is the Griffiths–Engen–McCloskey distribution, and is a function of X_i. The method is computationally expensive, and at present, an implementation only exists for $p \leq 2$. We compare our method to the method of Arbel *and others* (2016) with respect to both estimation error and computation time. We also note the related method of Ren *and others* (2017), which also implements a nonparametric model for W given Z but with a marginal prior distribution for Z_i given by a Gamma process. However, since it is unable to handle continuous covariates, we do not consider it further.

There exist many statistical models for species counts. However, most of these models do not model species relative abundances, and so cannot directly be used for estimating diversity indices that are functions of relative abundance. For example, the classical model of Dorazio and Royle (2005) models the presence and detection probabilities of species, but not species relative abundances (similar for Yamaura *and others*, 2011 but using presence data, rather than count data). Similarly, Hui *and others* (2015) (see also Letten *and others*, 2015) propose a latent variable model for species counts, and Pollock *and others* (2014) propose a model for species presence, but neither model estimates latent relative abundance. It has been previously noted (Gloor *and others*, 2016, 2017) that modeling relative abundance data with non-compositional models can lead to incorrect conclusions because the unit-sum constraint can alter the apparent direction of changes to the community. For example, if taxa 1, 2, and 3 exist in absolute abundance of, respectively, 100 units, 20 units, and 20 units before a treatment and 100 units, 40 units, and 20 units after a treatment, the relative abundance of taxa 1 and 3 have decreased, even though their absolute abundance was unchanged by the treatment. The model that we propose in Section 3 explicitly accounts for the compositional nature of microbiome data.

2.2. *β-Diversity*

Similar to α-diversity, a large number of different β-diversity metrics exist, each highlighting different features of differences in communities. Legendre and Legendre (2012, Table 7.2) provide a list of 26 β-diversity metrics along with some discussion. However, in comparison to α-diversity estimands, there exists almost no statistical literature on estimating β-diversity indices: estimating β-diversity indices is almost exclusively performed using plug-in estimators.

In general, small values of a β-diversity index indicate that the communities have similar compositions, while large values indicate that the relative abundances differ between communities, or that few taxa are shared by the communities. This interpretation holds for both the Bray–Curtis and Euclidean indices discussed below.

2.2.1. *Bray–Curtis dissimilarity* The (observed) Bray–Curtis index (Bray and Curtis, 1957) is defined as

$$\hat{\beta}_{ij,BC,\text{plug-in}} = 1 - 2\frac{\sum_{q\in\mathcal{C}_i\cup\mathcal{C}_j}\min(W_{iq}, W_{jq})}{M_i + M_j}. \tag{2.3}$$

While we have not found any discussion of the target estimand in the literature, (2.3) suggests that $\beta_{ij,BC} = 1 - \sum_{q\in\mathcal{C}}\min(Z_{iq}, Z_{jq})$ is the target estimand. Interestingly, in contrast to the other β-diversity indices discussed in the section, this estimate is not the MLE under a multinomial model.

While Arbel *and others* (2016) focused on estimating α-diversity, because their method estimates the latent composition matrix Z, we also compare our proposed method to the estimate $1 - \sum_{q\in\mathcal{C}}\min\left(\hat{Z}_{iq}^{(\text{Arbel})}, \hat{Z}_{jq}^{(\text{Arbel})}\right)$, where $\hat{Z}^{(\text{Arbel})}$ is the latent composition matrix estimate based on the procedure of Arbel *and others* (2016).

2.2.2. *Euclidean distance* Finally, we mention the Euclidean distance between the relative abundance vectors, $\beta_{ij,\text{ED}} = \sqrt{\sum_{q\in\mathcal{C}}(Z_{iq} - Z_{jq})^2}$, whose plug-in estimate is $\hat{\beta}_{ij,\text{ED}} = \sqrt{\sum_{q\in\mathcal{C}_i\cup\mathcal{C}_j}\left(\frac{W_{iq}}{M_i} - \frac{W_{jq}}{M_j}\right)^2}$. We are not aware of any other estimates for the Euclidean distance between relative abundances in the literature, but we will also compare to the estimate $\hat{\beta}_{ij,\text{ED,Arbel}} = \sqrt{\sum_{q\in\mathcal{C}}(\hat{Z}_{iq}^{(\text{Arbel})} - \hat{Z}_{jq}^{(\text{Arbel})})^2}$.

3. Estimating diversity in networked compositional data

Members of ecological communities interact, displaying repeatable patterns in many different environmental settings (Faust and Raes, 2012). For example, organisms may compete for resources, prey on each other, or cooperate in a symbiotic relationship. In the last decade, many methods have been developed to estimate the co-occurrence patterns of ecological communities, such as SparCC (Friedman and Alm, 2012) and SPIEC-EASI (Kurtz *and others*, 2015). We will refer to co-occurrence patterns as *ecological networks*. As we show under simulation, ecological networks can have substantial effects on estimates of diversity. Here, we propose an approach to estimating diversity in the presence of an ecological network.

3.1. *Compositional data models*

While the multinomial distribution is the canonical model for compositional data, the covariance between the number of observations in different categories is constrained to be negative. To deal with this issue, Aitchison (1982, 1986) developed the log-ratio model. This models the counts W_{iq} as independent draws from a multinomial distribution,

$$p(W|Z) \propto \prod_{i=1}^{n}\prod_{q=1}^{Q} Z_{iq}^{W_{iq}}, \tag{3.1}$$

where $Z \in \mathbb{R}^{n\times Q}$ is a matrix-valued latent random variable that gives the underlying composition matrix for each of the communities: $\sum_{q=1}^{Q} Z_{iq} = 1$ for all i. It then employs the log-ratio transformation by fixing a "baseline" taxon (taxon D) for comparison:

$$Y_{iq} = \phi(Z_{iq}) = \left\{ \log\left(\frac{Z_{iq}}{Z_{iD}}\right) \right\}_{q=1,\ldots,D-1,D+1,\ldots Q}. \tag{3.2}$$

Note that the log-ratio transformation $\phi : \mathbb{R}^{Q} \to \mathbb{R}^{Q-1}$ is invertible with inverse ϕ^{-1}:

$$Z_{iq} = \phi^{-1}(Y_{iq}) := \left\{ \begin{array}{ll} \frac{\exp(Y_{iq})}{\sum_{q\neq D}\exp(Y_{iq})+1} & q \neq D \\ \frac{1}{\sum_{q\neq D}\exp(Y_{iq})+1} & q = D \end{array} \right\}.$$

To permit flexible co-occurrence structures between the taxa, the log-ratios are modeled by a multivariate normal distribution:

$$f(\mathbf{Y}_i|\mu,\Sigma) \propto |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2}(\mathbf{Y}_i - \mu_i)^T \Sigma^{-1}(\mathbf{Y}_i - \mu_i) \right\}. \tag{3.3}$$

Finally, the mean of $\mathbf{Y}_i$ is linked to covariates via $\mu_i = X_i^T\gamma$, where $\gamma \in \mathbb{R}^{p\times(Q-1)}$. Under this model, γ_{rq} gives the expected increase in $\log\left(\frac{Z_{iq}}{Z_{iD}}\right)$ for a one-unit increase in X_{ir}. For a discussion of the interpretation of this model on the scale of Z_{iq}, we refer the reader to Billheimer *and others* (2001).

This model does not impose that all communities with the same covariate vector X_i have the same latent relative abundance vector $\mathbf{Z}_i$. However, under this model, the expectation of $\phi(\mathbf{Z}_i)$ is the same for all communities with the same covariate vector. Therefore, communities with the same environmental conditions are not constrained to have the same diversity under our model. We also note that this model is predicated on the assumption that the counts are conditionally independent given the covariate matrix $X \in \mathbb{R}^{n\times p}$. Therefore, this model does not apply to spatially or temporally correlated data. We analyze the effect of temporal dependence in Section S2 of the supplementary material available at *Biostatistics* online.

3.2. *Estimating diversity in the presence of a network*

We propose using the log-ratio model to estimate α-diversity and β-diversity. To our knowledge, this is the first proposal to estimate these diversity parameters under a model that explicitly models taxon–taxon co-occurrence structure. Let $\hat{\gamma}$ be an estimate of γ under the log-ratio model. We take a frequentist approach to estimation, and discuss maximum likelihood estimators in Section 4.1 and penalized maximum likelihood estimators in Section 4.2.

Suppose that we wish to estimate the α-diversity of a community with covariate vector $X_i \in \mathbb{R}^p$. Define $\hat{Y}_i = X_i^T\hat{\gamma}$, the expected value of the random variable $\mathbf{Y}_i$, and define $\hat{Z}_i = \phi^{-1}(\mathbf{Y}_i)$, the fitted value of the latent composition.

We then propose the following estimate of any α-diversity index $f : \mathbb{S}^{C-1} \to \mathbb{R}$:

$$\hat{\alpha}_i = f(\hat{Z}_i). \tag{3.4}$$

More explicitly, $\hat{\alpha}_{i,\text{Sh,proposed}} = -\sum_q \hat{Z}_{iq}\log\hat{Z}_{iq}$ and $\hat{\alpha}_{i,\text{Si,proposed}} = \sum_q(\hat{Z}_{iq})^2$ give our proposed estimates of the Shannon and Simpson indices. Similarly, for any β-diversity index $g : \mathbb{S}^{C-1}\times\mathbb{S}^{C-1} \to \mathbb{R}$, we propose

$$\hat{\beta}_{ij} = g(\hat{Z}_i, \hat{Z}_j), \tag{3.5}$$

such as $\hat{\beta}_{ij,\mathrm{BC,proposed}} = 1 - \sum_q \min(\hat{Z}_{iq}, \hat{Z}_{jq})$ and $\hat{\beta}_{ij,\mathrm{ED,proposed}} = \sqrt{\sum_q (\hat{Z}_{iq} - \hat{Z}_{jq})^2}$ for the Bray–Curtis and Euclidean diversity indices. Note that if $\hat{\gamma}$ is the MLE of γ, then by invariance, the proposed estimates are the MLEs of the diversity indices.

This approach to diversity estimation has a number of key advantages not shared by other methods. Fundamentally, rather than describing a quantity associated with the sample (as is the case with plug-in estimates), the estimand is the diversity of the community from which the sample was drawn. This means that information is shared across all samples to obtain more precise and accurate estimates. Furthermore, we can use the model to estimate the diversity of communities for which ecosystem survey data are not available but for which covariate information exists. While these advantages are shared with the method of Arbel *and others* (2016), our method is fast and is available as an open-source R package with examples and tutorials illustrating its use.

4. Parameter estimation

4.1. *Estimating model parameters*

To estimate the parameter set $\eta = (\gamma, \Sigma)$, we take a frequentist approach via maximum likelihood. If Y were known, our optimization problem would be to find

$$\hat{\eta} = \underset{\eta}{\operatorname{argmax}} \sum_{i=1}^{n} [\log Pr(\mathbf{W}_i|\mathbf{Y}_i) + \log f(\mathbf{Y}_i|\eta)], \tag{4.1}$$

where

$$\log Pr(\mathbf{W}_i|\mathbf{Y}_i) = \sum_{q \neq D} W_{iq} Y_{iq} - M_i \log \left(\sum_{q \neq D} \exp(Y_{im}) + 1 \right) \tag{4.2}$$

and

$$\log f(\mathbf{Y}_i|\eta) = -\frac{1}{2} \log(|\Sigma|) - \frac{1}{2} (\mathbf{Y}_i - \mu_i) \Sigma^{-1} (\mathbf{Y}_i - \mu_i). \tag{4.3}$$

Alas, since Y is a latent random variable, we cannot directly optimize (4.1). Instead, we use the Expectation–Maximization (EM) algorithm (Dempster *and others*, 1977). The expected complete log-likelihood is

$$Q(\eta|\eta^{(t)}) = -\frac{n}{2} \log(|\Sigma|) - \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}_{Y|(W,\eta^{(t)})} [(\mathbf{Y}_i - \mu_i)^T \Sigma^{-1} (\mathbf{Y}_i - \mu_i)]. \tag{4.4}$$

To estimate this expectation numerically, we follow Xia *and others* (2013) and use the Metropolis–Hastings (MH) algorithm. Let $\left\{ \mathbf{Y}_i^{(r)} \right\}_{r=1}^{R}$ be R draws from the distribution of $\mathbf{Y}_i|\mathbf{W}_i, \eta^{(t)}$. Given these draws, we can approximate the expectation as follows:

$$\mathbb{E}_{Y|(W,\eta^{(t)})} [(\mathbf{Y}_i - \mu_i)^T \Sigma^{-1} (\mathbf{Y}_i - \mu_i)] \approx \frac{1}{R} \sum_{r=1}^{R} (\mathbf{Y}_i^{(r)} - \mu_i^{(t)})^T (\Sigma^{(t)})^{\dagger} (\mathbf{Y}_i - \mu_i^{(t)}), \tag{4.5}$$

where † is the generalized inverse.

To generate the rth draw from $f(\mathbf{Y}_i|\mathbf{W}_i, \eta^{(t)})$, we simulate a proposal $\mathbf{Y}_i^{(*)} \sim \mathcal{N}_{Q-1}(\mathbf{Y}_i^{(r-1)}, vI_{Q-1})$, where v is a tuning parameter controlling the step size and I_{Q-1} is the identity matrix of dimension $Q-1$. We then calculate the Metropolis acceptance ratio

$$r(\mathbf{Y}_i^{(*)}|\mathbf{Y}_i^{(r-1)}) = \min\left\{1, \frac{f(\mathbf{Y}_i^{(*)}|\mathbf{W}_i, \eta^{(t)})}{f(\mathbf{Y}_i^{(r-1)}|\mathbf{W}_i, \eta^{(t)})}\right\},$$

and simulate $u \sim \text{Uniform}(0,1)$. We set $\mathbf{Y}_i^{(r)} = \mathbf{Y}_i^{(*)}$ if $u \leq r(\mathbf{Y}_i^{(*)}|\mathbf{Y}_i^{(r-1)})$, otherwise, we set $\mathbf{Y}_i^{(r)} = \mathbf{Y}_i^{(r-1)}$. By initializing $\mathbf{Y}_i^{(0)} = \phi\left(\frac{\mathbf{W}_i}{M_i}\right)$, setting $v = 0.01$, and discarding the first 500 draws, we observe convergence to the target distribution on a variety of microbiome datasets, and acceptance ratios ranging 30–40%.

Having obtained an estimate of the expectation in (4.4), we turn our attention to maximizing $Q(\eta|\eta^{(t-1)})$. Define $\eta^{(t)} = \text{argmax}_\eta\, Q(\eta|\eta^{(t-1)})$. Given our draws $\left\{\mathbf{Y}_i^{(r)}\right\}_{r=1}^R$ from $f(\mathbf{Y}_i|\mathbf{W}_i, \eta^{(t)})$, our M-step of the EM algorithm gives the following estimates:

$$\gamma^{(t+1)} = \frac{1}{R}\sum_{r=1}^{R}(X^TX)^\dagger X^T Y^{(r)}, \tag{4.6}$$

$$\mu_i^{(t+1)} = X_i^T\gamma^{(t+1)}, \tag{4.7}$$

$$\Sigma^{(t+1)} = \frac{1}{nR}\sum_{r=1}^{R}\sum_{i=1}^{n}\left\{\mathbf{Y}_i^{(r)} - \mu_i^{(t)}\right\}\left\{\mathbf{Y}_i^{(r)} - \mu_i^{(t)}\right\}^T, \tag{4.8}$$

where $X \in \mathbb{R}^{n\times p}$ and $Y^{(r)} = \left(\mathbf{Y}_1^{(r)}, \ldots, Y_n^{(r)}\right)^T \in \mathbb{R}^{n\times(Q-1)}$. Inspection of convergence diagnostics (such as trace plots) on a variety of datasets indicates that $R = 500$ and $\hat{\eta} = \eta^{(t)}$ for $t = 6$ is generally sufficient to achieve stable estimates (see Section S4 of the supplementary material available at *Biostatistics* online). We run the MH algorithm to approximate the distribution of $\mathbf{Y}_i|\mathbf{W}_i, \eta^{(t)}$ in parallel over $i = 1, \ldots, n$ to reduce computation time. Our code is publicly available as an `R` package and can be found at github.com/adw96/DivNet.

4.2. *Variance estimation*

To test hypotheses about changes in diversity over environmental gradients, it is necessary to have accurate estimates of the variance of the diversity estimates. These variance estimates can then be used in hypothesis testing (e.g., using the method of Willis *and others* (2016)). We consider both parametric and nonparametric bootstrap approaches to estimating the variance of the diversity estimates produced by our model and evaluate them under simulation. For a given dataset (W, X), let $\hat{\gamma}$ and $\hat{\Sigma}$ be the estimated values of γ and Σ estimated by the algorithm described in Section 4.1.

The parametric bootstrap approach to estimating $\text{Var}(\hat{\alpha}_i)$ and $\text{Var}(\hat{\beta}_{ij})$ for arbitrary diversity indices works as follows: B datasets are simulated from the log-ratio model with $\mu = X\hat{\gamma}$ and $\Sigma = \hat{\Sigma}$. Then, for each of the B simulated datasets, bootstrap estimates $\{(\hat{\gamma}^{(b)}, \hat{\Sigma}^{(b)})\}_{b=1}^B$ are obtained using the algorithm described in Section 4.1, and an estimate of the diversity index for community i is obtained based on each simulated dataset (i.e., $\{\hat{\alpha}_i^{(b)}\}_{b=1}^B$). The parametric bootstrap estimate of $\text{Var}(\hat{\alpha}_i)$ is then $\widehat{\text{Var}}_b(\hat{\alpha}_i^{(b)})$, where $\widehat{\text{Var}}(\cdot)$ is the sample variance. An estimate of the variance of any β-diversity index can be obtained in the same way.

We also consider a nonparametric bootstrap approach to estimating the variance of our estimates. We investigate the nonparametric bootstrap for completeness. To construct a nonparametric bootstrap estimate, we uniformly at random select with replacement n_{sub} elements from $\{1,\dots,n\}$ to obtain a set which we call $\mathcal{B}$. We then estimate $(\hat{\gamma}^{(\mathcal{B})}, \hat{\Sigma}^{(\mathcal{B})})$ from $(W^{(\mathcal{B})}, X^{(\mathcal{B})})$, where $W^{(\mathcal{B})}$ and $X^{(\mathcal{B})}$ are the rows of W and X with row index in $\mathcal{B}$, and use $\{(\hat{\gamma}^{(\mathcal{B})}, \hat{\Sigma}^{(\mathcal{B})})\}$ estimates to obtain $\hat{\alpha}_i^{(\mathcal{B})}$. We repeat this process B times to obtain a set of estimates $\{\hat{\alpha}_i^{(\mathcal{B}_b)}\}_{b=1}^{B}$ from which we calculate the nonparametric bootstrap estimate $\widehat{\text{Var}}(\hat{\alpha}_i) = \widehat{\text{Var}}_b(\hat{\alpha}_i^{(\mathcal{B}_b)})$ (and similarly for β-diversity).

The parameter Σ drives the variance in the log-ratio model: as $||\Sigma||_\infty \to 0$, the distribution of W converges to a multinomial distribution. Therefore, the overdispersion of the log-ratio model relative to the multinomial model is driven by Σ. However, the number of taxa often greatly exceeds the number of communities obtained in microbiome surveys, and in this setting, $(\Sigma^{(t)})^\dagger$ may be a poor estimate of Σ^{-1} in (4.5), even for large t. We therefore consider replacing $(\Sigma^{(t)})^\dagger$ in (4.5) with a regularized estimate obtained from the graphical lasso (Friedman *and others*, 2008; Witten *and others*, 2011). Following the popular microbial network estimation software SPIEC-EASI (Kurtz *and others*, 2015), we use stability selection to select the regularization parameter (Liu *and others*, 2010; Kurtz *and others*, 2015). We also consider replacing $(\Sigma^{(t)})^\dagger$ with the MLE restricted to the class of diagonal covariance matrices. Note that this approach to covariance estimation ignores variance attributable to inter-taxon interactions but allows for overdispersion relative to the multinomial due to within-taxon interactions.

We evaluate the performance of these 6 approaches to estimating the variance of diversity indices (two approaches to estimating the variance for each of three approaches to estimating the inverse covariance) under simulation. We design our simulation to mimic the dataset analyzed in Section 6, but with varying Q, the number of taxa and the size of the covariance matrix to be estimated. As is the case for the dataset of Section 6, we fix $p = 2$, $n = 12$, and set $X = (\mathbf{1}_n^T, (\mathbf{0}_{2n/3}, \mathbf{1}_{1n/3})^T)$. Note that our method can accommodate both discrete and continuous covariates, but we choose discrete covariates for all simulation studies to reflect the structure of the dataset analyzed in Section 6. Let $\mathcal{W}^Q$ be the columns of the count matrix W of Section 6 corresponding to the Q most common taxa over all samples. Let $\mathbf{Y}_i^Q = \phi(\mathcal{W}_i^Q) \in \mathbb{R}^{Q-1}$, and $Y^Q = [\mathbf{Y}_1^Q \cdots \mathbf{Y}_n^Q] \in \mathbb{R}^{n\times(Q-1)}$. We set $\gamma^Q = (X^TX)^{-1}X^TY^Q$ and Σ^Q to be the covariance of the columns of $Y^Q - X\gamma^Q$, and for each Q, we simulate data according to the log-ratio model with parameters γ^Q, Σ^Q and $M_i = \sum_q W_{iq}$. Specifically, to simulate from the log-ratio model with parameters (γ, Σ, X, M), we first simulate a matrix $Y \in \mathbb{R}^{n\times(Q-1)}$ with ith row $\mathbf{Y}_i \sim \mathcal{N}(X_i^T\gamma, \Sigma)$, then calculate the matrix Z with ith row $\mathbf{Z}_i = \phi^{-1}(\mathbf{Y}_i)$ (see (3.2)), and finally simulate the matrix $W \in \mathbb{Z}^{n\times Q}$ with $\mathbf{W}_i \sim \text{Multinomial}(M_i, \mathbf{Z}_i)$. Noting that n is small at $n = 12$ (as is often the case for microbiome analyses), we choose $B = 3$ simulated datasets for the parametric bootstrap and $B = 3$ subsamples of size $n_{\text{sub}} = 6$ for the nonparametric bootstrap approach, but to ensure that our simulation results are accurate we average over 25 simulation replicates.

We compare the estimated variance of the six methods in Figure 1 for a varying number of taxa Q. Only the variance of the Shannon index and Bray–Curtis index are shown, but similar patterns were observed for all indices. We observe that both parametric and nonparametric bootstrap variances are of similar magnitude, with parametric approaches generally having slightly lower median variance (left panels). In addition, to confirm that the estimated variance does not underestimate the true variance, we compare the difference between the estimated variance and the true variance for each method (right panels). The true variance of each method is estimated by repeatedly simulating data according to (γ^Q, Σ^Q, M), estimating the diversity index for each simulated dataset and each covariance estimate, and calculating the variance of the estimated indices. We observe that the median difference between the true variance and the stated variance is near zero for the parametric approaches, but negative for the nonparametric approaches, indicating that nonparametric approaches tend to underestimate the true variance. However, none of the three approaches to covariance estimation show substantial advantage over the others. This

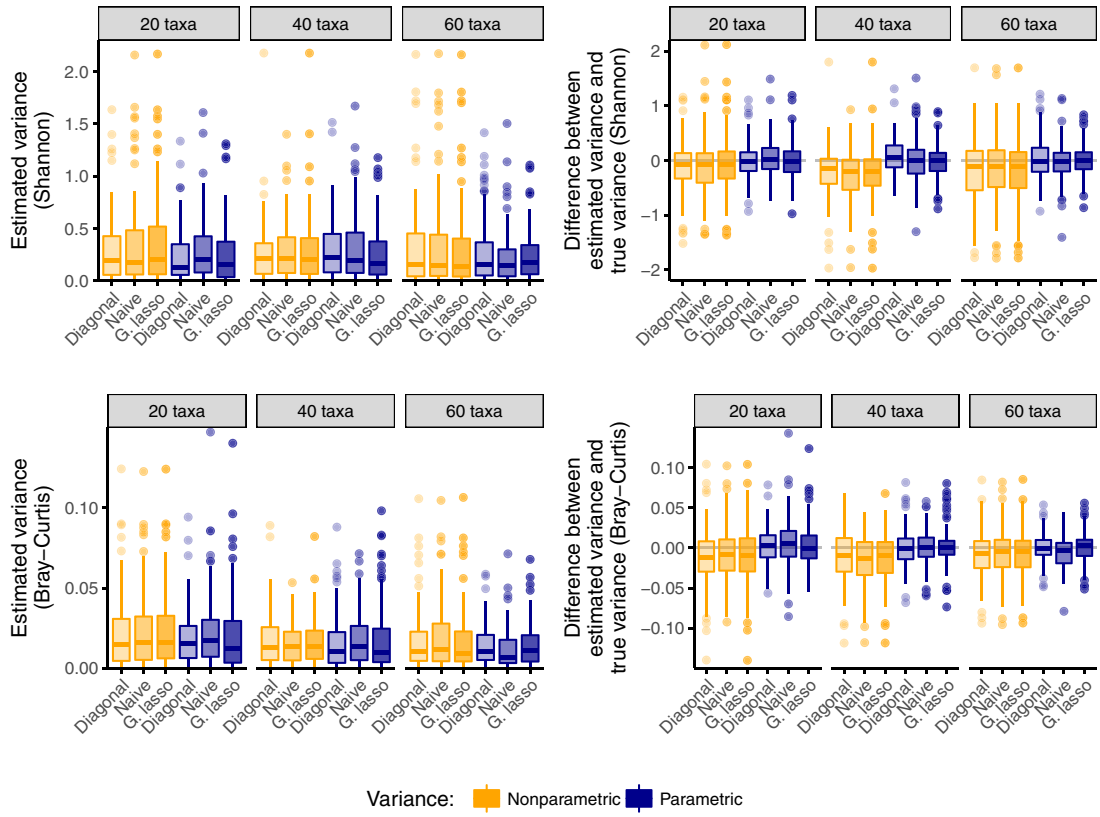

Fig. 1. A comparison of nonparametric and parametric bootstrap approaches to estimating the variance of diversity estimates under a model that incorporates microbial co-occurrence patterns. The parametric bootstrap has lower variance than the nonparametric bootstrap (left panel), and the median difference with true variance close to zero (right panel). No approach to covariance estimation consistently outperforms other approaches.

suggests that the primary driver of variance in estimating diversity in microbial communities is within-taxon interactions (the diagonal elements of Σ), rather than between-taxon interactions (the off-diagonal elements of Σ). Given these results, we select the naïve (generalized inverse of the sample covariance) approach to estimating $(\Sigma^{(t)})^{-1}$ as our default method. This approach is less computationally expensive than fitting the graphical lasso, while still permitting between-taxon interactions in the model. However, the functionality to estimate Σ via a structured approach is implemented in our `R` package.

5. SIMULATION STUDY

An investigation of the performance of our proposed method is available as supplementary material available at *Biostatistics* online. We investigated the performance of the method when data are generated according to the model described in Section 3.1 (Section S1 of the supplementary material available at *Biostatistics* online), when the data are generated according to the stochastically perturbed discrete-time Lotka–Volterra (LV) model of Fisher and Mehta (2014) (Section S2 of the supplementary material available at *Biostatistics* online), and when data are generated according to a nonlinear model on the log-ratio scale (Section S3 of the supplementary material available at *Biostatistics* online). In Section S1

of the supplementary material available at *Biostatistics* online, we investigated the effect of sample size (Section S1.1 of the supplementary material available at *Biostatistics* online), co-occurrence structure (Section S1.2 of the supplementary material available at *Biostatistics* online), and number of taxa (Section S1.3 of the supplementary material available at *Biostatistics* online). In Section S2 of the supplementary material available at *Biostatistics* online, we investigated the effect of number of taxa and number of time points. In Section S3 of the supplementary material available at *Biostatistics* online, we investigated both a quadratic and an exponential trend and varied the degree of curvature for each. We found that the proposed method strongly outperforms competitors when data are generated according to the model described in Section 3.1. When data are generated according to the stochastically perturbed discrete-time LV model, its performance suffers, especially when there are a large number of taxa and a small number of time points. However, estimation is relatively robust to nonlinear trends. We refer the reader to supplementary material available at *Biostatistics* online for details on the data generating processes and our results.

6. Data analysis: seafloor microbial diversity

Because of its coarse nature as a community-level summary, diversity analyses are especially relevant to studies of novel ecological communities. Lee *and others* (2015) collected and analyzed microbial communities living on seafloor rocks on the Dorado Outcrop, an area of exposed basalt on the East Pacific Rise. Hydrothermal vents such as the Dorado Outcrop inform our understanding of microbe–mineral interactions in the subsurface. Samples were collected from the seafloor rock, including glassy, altered basalts ("glassy," $n = 4$) and highly altered basalts ("altered," $n = 8$). Analysis of the microbial communities on these rocks revealed 1425 distinct microbial taxa in glassy and altered basalts after filtering for low quality sequences (see Lee *and others* (2015) and Lee (2018) for details surrounding sequencing and construction of the abundance table). Here, we investigate if the community-level structure differs between the different rock types.

We investigate 30 choices for the Qth taxon, whose abundance will be the denominator in the calculated log-ratios. Since $\frac{\partial}{\partial y}\log(x/y) = -1/y$ is smallest in absolute value for large y, we investigate the effect of setting Q to be a high abundance taxon. In particular, there were 86 amplicon sequence variants (ASVs) that were present in all samples, and so we uniformly at random select 10 ASVs from this collection of 86 ASVs, and compare the estimates of diversity obtained by setting each of these 10 taxa as the denominator taxon. We contrast these estimates with those obtained from ranging Q across the 10 most abundant taxa over all samples. We also compare 10 randomly selected taxa. The estimated Shannon, Simpson, Jaccard, and Euclidean diversities are shown in Figure 2 (2nd and 4th panels), indicating that, in practice, the diversity estimates are almost invariant to the choice of base taxon. We select Q to be ASV 2, (a Nitrospirae of order Nitrospirales), which was the most abundant taxon that was observed in every sample.

In contrast to the stability of diversity estimates with varying D, we find that the effect of perturbing the zero counts can be substantial (Figure 2, 1st and 3rd panels). As noted previously (Martín-Fernández *and others*, 2003; Cao *and others*, 2019a,b), W_{ij} is commonly zero for microbiome data, because many taxa do not occur in every sample (42% of the entries of our abundance table are zero). However $f(x, y) = \log(x/y)$ is only defined for $x, y > 0$, and so it is common to perturb the original abundance data W by adding a perturbation factor $\rho \in (0, 1)$ to create a new abundance table $W_{ij}^{(\rho)} = W_{ij} + \rho$, and the modeling the perturbed data $W^{(\rho)}$. In Figure 2, we observe sizeable changes in the diversity estimates when varying p close to zero (at most 26%, −50%, −24%, and −31% changes in Shannon, Simpson, Bray-Curtis, and Euclidean estimates for $\rho = 0.001$ compared to $\rho = 0.5$), but smaller changes when ρ is increased from 0.5 to 1 (at most 5%, −24%, −12%, and −13% changes for $\rho = 0.5$ to $\rho = 1$). We therefore follow Cao *and others* (2019a) and choose $p = 0.5$ as the perturbation parameter for the remainder of our analysis.

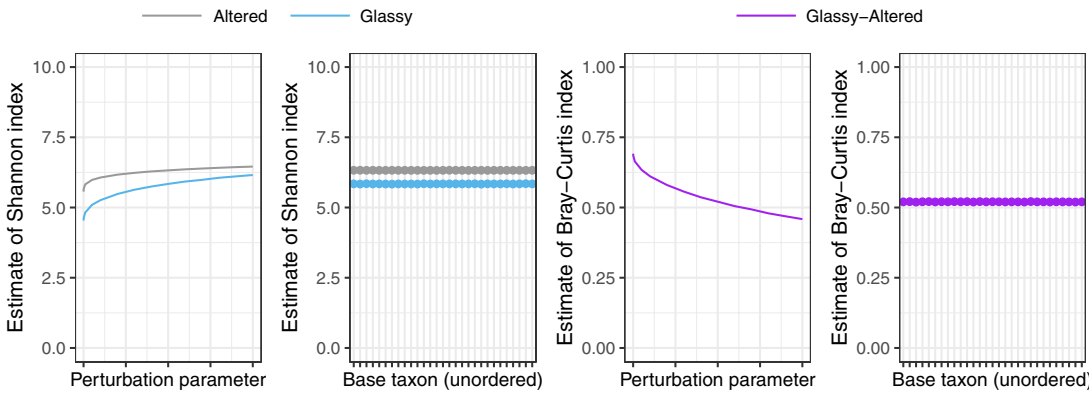

Fig. 2. The log-ratio model described in Section 3 can only be fit to data with a minimum abundance greater than zero. Abundance data for microbiome studies are generally sparse, and 42% of the observed abundances of the Lee *and others* (2015) dataset are zero. For this reason, it is common to add a perturbation offset ρ to the observed abundance table before fitting the log-ratio model. Here, we see that the estimated diversity does depend on the choice of ρ.

Throughout this article, we have argued that the multinomial model is misspecified for microbiome data. To investigate this claim for the dataset of Lee *and others* (2015), we fit the log-ratio model and calculate the eigenvalues of $\hat{\Sigma}$. Since the multinomial model is the limit of the model described in Section 3.1 as $\Sigma \to 0$, and the largest eigenvalue of $\hat{\Sigma}$ for this dataset is 422.87, this is strong evidence that the multinomial model is misspecified for this dataset.

Finally, we compare our estimates to the estimates obtained from other methods. Interval estimates are shown in Figure 3. The proposed method was fit in mode `tuning = "careful"`; the method of Arbel *and others* (2016) was run for 500 iterations, and convergence was confirmed via trace plots; and the method of Chao and Shen (2003) was run with the default $k = 40$ and 50 bootstrap resamples. Code for repeating the analysis is available at github.com/adw96/DivNet_supplementary.

While most methods produce similar estimates, we note a number of advantages of our proposal. Firstly, any diversity index that is a function of relative abundance can be estimated using our method, unlike the methods of Hsieh *and others* (2016) and Chao and Shen (2003). Secondly, our interval estimates are more symmetric around the median of the bootstrapped estimates compared to other estimates. Thirdly, while this analysis only included two covariates, our method can handle multiple covariates.

7. DISCUSSION

Despite substantial evidence that strong co-occurrence networks exist in microbial communities, and a growing body of literature concerned with estimating co-occurrence networks, no methods that explicitly incorporate co-occurrence networks into diversity estimation currently exist. Here we propose a new method, called `DivNet`, to fill this gap. `DivNet` is highly accurate when the log-ratio model is correctly specified, including when there are a large number of taxa. `DivNet` can be used to model count data arising from direct observations, flow cytometry, or high throughput sequencing technologies such as 16S amplicon sequencing. It is available as an open-source `R` package via github.com/adw96/DivNet.

By leveraging information from multiple samples, `DivNet` can estimate the relative abundance of a taxon in a community where it was not observed. However, a limitation of `DivNet` is that it does not estimate the number of taxa that were missing in all samples. Therefore, when there are a large number of latent taxa, `DivNet` may miss the effects of these low abundance taxa. This weakness is shared by the estimators of Arbel *and others* (2016) and Cao *and others* (2019b), while the estimators of

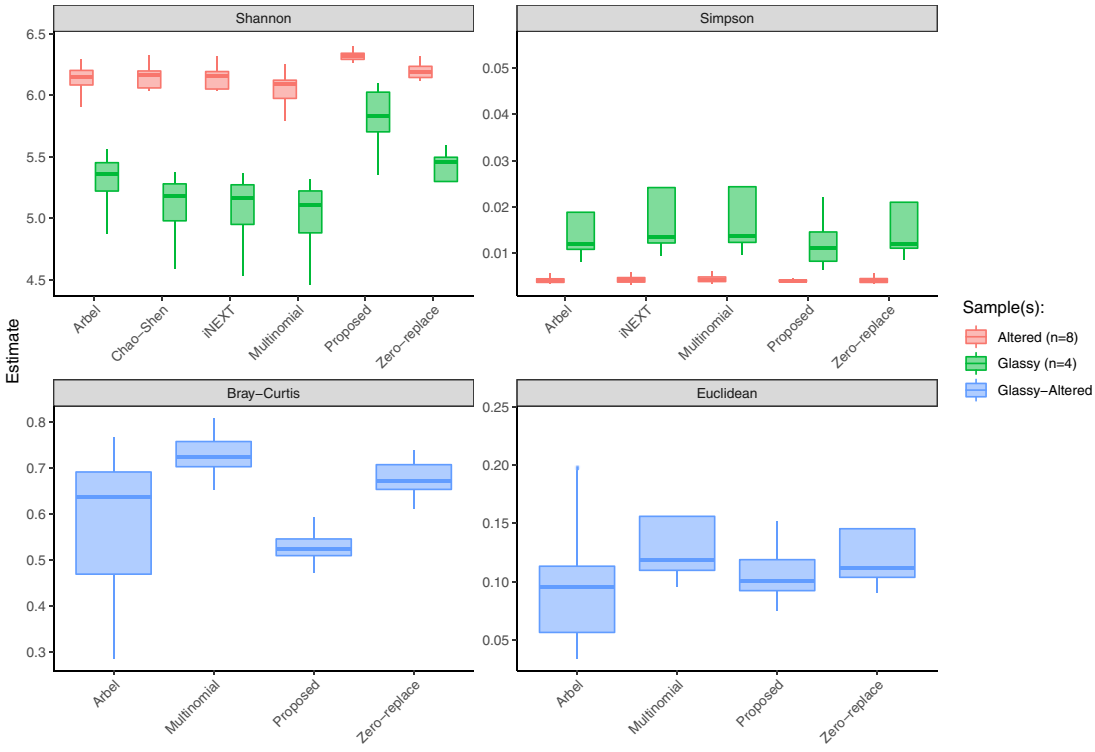

Fig. 3. Lee *and others* (2015) collected and analyzed microbial communities living on different types of seafloor basalts on the Dorado Outcrop. Here, we compare a variety of estimators for four diversity indices. 25% and 75% quantiles are shown.

Hsieh *and others* (2016) and Chao and Shen (2003) adjust for missing taxa (but are only applicable to α-diversity). However, the latter two estimators cannot handle covariates nor repeated samples, which contribute to the performance of our method. In the situation when no replicates or covariates are available, there are a large number of latent taxa, and β-diversity is not of interest, a practitioner may prefer these methods.

Under simulation, we demonstrated that `DivNet` performs favorably when data are generated independently and identically from a distribution where the taxa co-occur on the log-ratio relative abundance scale. This generally holds even when the set of covariates is misspecified, such as when there is an exponential trend on the log-ratio scale, but a linear or quadratic model is fit. We also found that the performance of `DivNet` suffers when data are generated according to a LV model. Under this model, the abundances of taxa are temporally correlated, and the co-occurrence network acts on the absolute, not relative, abundant scale. We found that for short LV-simulated time series with many taxa, other estimators may outperform `DivNet`. We note that other violations of conditional independence are likely to adversely affect `DivNet`'s performance, including spatial correlation. We encourage caution when applying `DivNet` to count data where observations are not independent. In practice, since the data generating process is generally not known, we recommend that the user contrast a number of different estimators before drawing conclusions about diversity.

We also note that it is common for ecologists to be interested in the ordering of diversity indices rather than their absolute values. We are not aware of a data analysis where the ordering of diversity across

a covariate has been altered by the choice to estimate diversity using `DivNet`. However, because the log-ratio model is overdispersed compared to a multinomial model, the standard errors of `DivNet` are larger than the standard errors for the MLE of a multinomial model, reflecting the additional uncertainty captured by the model.

We suggest four avenues for further research that would build upon our proposed method. The first is to construct an estimator under the log-ratio model that estimates the number of missing taxa. However, this would require a principled approach to estimating the ecological network of a taxon that was not observed in any sample. A second avenue for research is to impose some structure (e.g., sparsity) on the relative abundance parameter γ, whose dimension is large when there are a large number of taxa. Thirdly, since diversity indices that incorporate relative abundance and phylogenetic information are commonly used by ecologists, extending the method to incorporate phylogeny is an open problem. Finally, a generalization of the method that relaxes the assumption of independence to account for correlation between observations (e.g., due to spatial or temporal dependence) is yet to be developed, and is likely to outperform `DivNet` when observations are correlated.

The method described in this manuscript is available at github.com/adw96/DivNet. Code to reproduce the simulations, figures, and data analysis is available at github.com/adw96/DivNet_supplementary.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

We are grateful to the Associate Editor and two anonymous reviewers whose helpful suggestions greatly improved the manuscript. The authors are also grateful to Mike Lee for the dataset discussed in Section 6 and helpful discussions, to Daniela Witten for many insights on the model, and to Ali Shojaie and Erick Matsen for highlighting important references.

Conflict of Interest: None declared.

FUNDING

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R35 GM133420. The opinions expressed in this article are those of the authors and do not necessarily represent the official views of the NIGMS or the NIH.

REFERENCES

AITCHISON, J. (1982). The statistical analysis of compositional data. *Journal of Royal Statistical Society B Methodological* **44**, 139–160.

AITCHISON, J. (1986). The statistical analysis of compositional data, pp. 141–182. London: Chapman & Hall.

ARBEL, J., MENGERSEN, K. and ROUSSEAU, J. (2016). Bayesian nonparametric dependent model for partially replicated data: the influence of fuel spills on species diversity. *The Annals of Applied Statistics* **10**, 1496–1516.

BASHARIN, G. P. (1959). On a statistical estimate for the entropy of a sequence of independent random variables. *Theory of Probability and Its Applications* **4**, 333–336.

BILLHEIMER, D., GUTTORP, P. and FAGAN, W. F. (2001). Statistical interpretation of species composition *Journal of the American Statistical Association* **96**, 1205–1214.

BRAY, J. R. and CURTIS, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin *Ecological Monographs* **27**, 325–349.

Callahan, B. J., Sankaran, K., Fukuyama, J. A., McMurdie, P. J. and Holmes, S. P. (2016). Bioconductor workflow for microbiome data analysis: from raw reads to community analyses *F1000Research* **5**, 1492.

Cao, Y., Zhang, A. and Li, H. (2019a). Large covariance estimation for compositional data via composition-adjusted thresholding. *Journal of the American Statistical Association* **44**, 1–14.

Cao, Y., Zhang, A. and Li, H. (2019b). Multi-sample estimation of bacterial composition matrix in metagenomics data. *Biometrika*.

Chao, A. and Shen, T.-J. (2003). Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics* **10**, 429–443.

Chao, A., Wang, Y. T. and Jost, L. (2013). Entropy and the species accumulation curve: a novel entropy estimator via discovery rates of new species. *Methods in Ecology and Evolution* **4**, 1091–1100.

De'ath, G. (2012). The multinomial diversity model: linking Shannon diversity to multiple predictors. *Ecology* **93**, 2286–2296.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39**, 1–22.

Dorazio, R. M. and Royle, J. A. (2005). Estimating size and composition of biological communities by modeling the occurrence of species. *Journal of the American Statistical Association* **100**, 389–398.

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation* **61**, 1–10.

Faust, K. and Raes, J. (2012). Microbial interactions: from networks to models. *Nature Reviews Microbiology* **10**, 538.

Fisher, C. K. and Mehta, P. (2014). Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PLoS One* **9**, e102451.

Friedman, J. and Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Computational Biology* **8**, e1002687.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Frontiers in Microbiology* **8**, 57.

Gloor, G. B., Macklaim, J. M., Vu, M. and Fernandes, A. D. (2016). Compositional uncertainty should not be ignored in high-throughput sequencing data analysis. *Austrian Journal of Statistics* **45**, 73–87.

Hill, M. O. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology* **54**, 427–432.

Hsieh, T. C., Ma, K. H. and Chao, A. (2016). iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). *Methods in Ecology and Evolution* **7**, 1451–1456.

Hui, F. K., Taskinen, S., Pledger, S., Foster, S. D. and Warton, D. I. (2015). Model-based approaches to unconstrained ordination. *Methods in Ecology and Evolution* **6**, 399–411.

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J. and Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Computational Biology* **11**, 1–25.

Lee, M. (2018). Example marker-gene workflow. astrobiomike.github.io/amplicon/workflow_ex.

Lee, M. D., Walworth, N. G., Sylvan, J. B., Edwards, K. J. and Orcutt, B. N. (2015). Microbial communities on seafloor basalts at Dorado Outcrop reflect level of alteration and highlight global lithic clades. *Frontiers in Microbiology* **6**, 1470.

Legendre, P. and Legendre, L. F. J. (2012). *Numerical Ecology*, Vol. 24. Amsterdam: Elsevier.

LETTEN, A. D., KEITH, D. A., TOZER, M. G. and HUI, F. K. (2015). Fine-scale hydrological niche differentiation through the lens of multi-species co-occurrence models. *Journal of Ecology* **103**, 1264–1275.

LI, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application* **2**, 73–94.

LIU, H., ROEDER, K. and WASSERMAN, L. (2010). *Stability approach to regularization selection (stars) for high dimensional graphical models*, Advances in Neural Information Processing Systems. Vol. 24, pp. 1432–1440.

LOZUPONE, C. and KNIGHT, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology* **71**, 8228–8235.

MARTÍN-FERNÁNDEZ, J. A., BARCELÓ-VIDAL, C. and PAWLOWSKY-GLAHN, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology* **35**, 253–278.

MCCOY, C. O. and MATSEN, F. A. (2013). Abundance-weighted phylogenetic diversity measures distinguish microbial community states and are robust to sampling depth. *PeerJ* **1**, e157.

MILLER, G. A. (1955). Note on the bias of information estimates. *Information Theory in Psychology: Problems and Methods* **2**, 100.

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.

REN, B., BACALLADO, S., FAVARO, S., HOLMES, S. and TRIPPA, L. (2017). Bayesian nonparametric ordination for the analysis of microbial communities. *Journal of the American Statistical Association* **112**, 1430–1442.

SHANNON, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal* **27**, 379–423.

SIMPSON, E. H. (1949). Measurement of diversity. *Nature* **163**, 688.

VU, V. Q., YU, B. and KASS, R. E. (2007). Coverage-adjusted entropy estimation. *Statistics in Medicine* **26**, 4039–4060.

WEISS, S., XU, Z. Z., PEDDADA, S., AMIR, A., BITTINGER, K., GONZALEZ, A., LOZUPONE, C., ZANEVELD, J. R., V´AZQUEZ-BAEZA, Y., BIRMINGHAM, A. *and others* (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**, 27.

WILLIS, A. D., BUNGE, J. and WHITMAN, T. (2016). Improved detection of changes in species richness in high-diversity microbial communities. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **66**, 963–977.

WITTEN, D. M., FRIEDMAN, J. H. and SIMON, N. (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics* **20**, 892–900.

XIA, F., CHEN, J., FUNG, W. K. and LI, H. (2013). A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics* **69**, 1053–1063.

YAMAURA, Y., ANDREW ROYLE, J., KUBOI, K., TADA, T., IKENO, S. and MAKINO, S. (2011). Modelling community dynamics based on species-level abundance models from detection/nondetection data. *Journal of Applied Ecology* **48**, 67–75.

ZAHL, S. (1977). Jackknifing an index of diversity. *Ecology* **58**, 907–913.

ZHANG, Z. and ZHOU, J. (2010). Re-parameterization of multinomial distributions and diversity indices. *Journal of Statistical Planning and Inference* **140**, 1731–1738.

[*Received July 12, 2019; revised March 4, 2020; accepted for publication March 8, 2020*]