DivNet: Estimating diversity in networked communities

Amy D Willis and Bryan D Martin

Abstract: Diversity is a marker of ecosystem health in ecology, microbiology and immunology, with implications for disease diagnosis and infection resistance. However, accurately comparing diversity across environmental gradients is challenging, especially when number of different taxonomic groups in the community is large. Furthermore, existing approaches to estimating diversity do not perform well when the taxonomic groups in the community interact via an ecological network, such as by competing within their niche, or with mutualistic relationships. To address this, we propose DivNet, a method for estimating within- and between-community diversity in ecosystems where taxa interact via an ecological network. In particular, accounting for network structure permits more accurate estimates of *alpha*- and *beta*-diversity, even in settings with a large number of taxa and a small number of samples. DivNet is fast, accurate, precise, performs well with large numbers of taxa, and is robust to both weakly and strongly networked communities. We show that the advantages of incorporating taxon interactions into diversity estimation are especially clear in analyzing microbiomes and other high-diversity, strongly networked ecosystems. Therefore, to illustrate the method, we analyze the microbiome of seafloor basalts based on a 16S amplicon sequencing dataset with 1490 taxa and 13 samples.

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 samples were observed. 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 et al. 2016). As well as providing an indicator of human and environmental health in microbiology (Oakley et al. 2008, Lozupone et al. 2012), diversity metrics are also widely used in immunology (Gibson et al. 2009, Kaplinsky & Arnaout 2016) and information theory.

Consider a 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 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 (typically, the similarity between two communities' relative abundance vectors $z^{(1)}$ and $z^{(2)}$). β-diversity indices

summarize between-community structure, while α-diversity indices summarize within-community structure. Specific examples of α- and β-diversity indices 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 & Zhou 2010, Hsieh et al. 2016, Cao et al. 2017). Fortunately, there exist sophisticated models for community composition data that permit more a flexible co-occurrence structure than that implied by the multinomial distribution. In this paper, we use models that explicitly permit co-occurrence of taxa (commonly referred to as ecological networks) to estimate community-level diversity.

In addition to incorporating network structure, the proposed method has a number of advantages over existing methods for diversity estimation and diversity-related hypothesis testing. Most notably, while almost all existing methodology for estimating diversity either estimates the diversity of each sample (for α-diversity) or pairs of samples (for β-diversity), our method pools information across multiple samples to estimate the diversity of the ecological communities from which the samples were drawn. Therefore, rather than estimating the diversity of a sample based only on abundance information obtained from that sample, abundance information from all samples is used to improve diversity estimation. This methodology also permits a principled method for predicting diversity in ecosystems that were not sampled. Our method achieves substantial improvements in estimation performance. The method, called `DivNet`, is available as a `R` package via `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 estimates. The performance of the method is evaluated in Section 5, before an example of the method is discussed in Section 6. We conclude with a discussion of the method and avenues for future research in Section 7.

2. Literature review: Estimating α- and β-diversity

Suppose that we have samples from $i = 1, \ldots, n$ ecosystems. Let $\mathcal{C}_i$ denote the set of all taxa in ecosystem i, and let $C_i = |\mathcal{C}_i|$ denote the number of taxa in the ith ecosystem. Let $\mathcal{C} = \cup_i \mathcal{C}_i$, and let $Q = |\mathcal{C}|$ denote the number of species present in one or more ecosystems. Finally, let $q = 1, \ldots, Q$ index the Q taxa. While not all taxa must be present in all ecosystems, we construct this set to ensure that the indexing is consistent. We impose the restriction that Q is known (see Section 7 for a discussion). Let $Z_{iq} \in [0, 1]$ denote the (unknown) relative abundance of taxon q in ecosystem i, noting that $\sum_{q=1}^{Q} Z_{iq} = 1$. Associated with each ecosystem is a known vector of covariates $X_i \in \mathbb{R}^p$.

Suppose that from the ith ecosystem, M_i individuals are observed and clas-

sified into the q taxonomic groups. Let W_{iq} denote the number of times that taxon q was observed in sample i. Therefore, to estimate summary statistics associated with the i ecosystems, 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 paper 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 & Matsen 2013) or UniFrac (Lozupone & 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 ecosystems. Two of the most common indices are the Shannon entropy (also called the Shannon index), and the Simpson index. The Shannon index places more emphasis on rare species than the Simpson index (because $-x\log x > x^2$ for x close to zero; see Eqs. (1) and (5)). Therefore, in ecosystems where rare species are significant drivers of ecosystem health, the Shannon index may be preferred over the Simpson index (for example, see Oakley et al. (2008)). While the diversity estimation framework that we will introduce is applicable to any α-diversity index that is a function of taxon abundance, 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 ecosystem i is defined as

$$\alpha_{i,Sh} = -\sum_{q\in\mathcal{C}_i} Z_{iq}\log(Z_{iq}). \tag{1}$$

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

Under the model $\mathbf{W}_i \sim Multinomial(M_i, \mathbf{Z}_i)$, the maximum likelihood estimate (MLE) of $\alpha_{i,Shannon}$ is

$$\hat{\alpha}_{i,Sh,plug-in} = -\sum_{q\in\mathcal{C}_i} \frac{W_{iq}}{M_i}\log\left(\frac{W_{iq}}{M_i}\right), \tag{2}$$

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 et al. 2017, Willis 2017). The multinomial MLE is often referred to as the *plug-in* estimate (Vu et al. 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 & Shen (2003) proposed using the Good-Turing estimate of species richness and adjusting for the missing taxa, obtaining the estimate

$$\hat{\alpha}_{i,Sh,Chao-Shen} = -\sum_{q \in \mathcal{C}_i} \frac{\hat{C}_i \hat{\pi}_{iq} \log(\hat{C}_i \hat{\pi}_{iq})}{1 - (1 - \hat{C}_i \hat{\pi}_{iq})^n}, \tag{3}$$

where $\hat{\pi}_{iq} = W_{iq}/M_i$ and $\hat{C}_i = 1 - \sum_q \mathbb{1}_{\{W_{iq}=1\}} / \sum_q W_{iq}$. Vu et al. (2007) show that this estimator is consistent and converges with the optimal rate $O_P(1/\log(M_i))$.

More recently, Chao et al. (2013) proposed to correct bias due to latent taxa by subsampling taxa and extrapolating from the sequentially smaller subsamples. The method is implemented in the R package `iNEXT` (Hsieh et al. 2016), against which we compare our method. We note that the subsampling procedure of `iNEXT` involves subsampling the taxa independently, which reflects 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 et al. (2003) propose replacing observed values of W_{ij} that are exactly zero with 0.5, and so Cao et al. (2017) consider the resulting *zero-replace* α-diversity estimator

$$\hat{\alpha}_{i,Sh,ZR} = -\sum_{q \in \mathcal{C}} \frac{W_{iq} \vee 0.5}{\sum_{r \in \mathcal{C}} W_{ir} \vee 0.5} \log\left(\frac{W_{iq} \vee 0.5}{\sum_{r \in \mathcal{C}} W_{ir} \vee 0.5}\right), \tag{4}$$

and also extend this idea to imputing zero elements of W via a low-rank matrix projection using a regularization approach based on 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*:

$$\alpha_{i,Si} = \sum_{q \in \mathcal{C}_i} Z_{iq}^2. \tag{5}$$

Similar to the Shannon index, the most common estimate of the Simpson index is the plug-in estimate:

$$\hat{\alpha}_{i,Si,plug-in} = \sum_{q \in \mathcal{C}_i} \left(\frac{W_{iq}}{M_i}\right)^2. \tag{6}$$

In comparison with Shannon entropy estimation, research concerning optimality of estimates of the Simpson index is relatively recent. Zhang & Zhou (2010) demonstrated that under independent sampling from a multinomial distribution,

$$\hat{\alpha}_{i,Si,Zhang-Zhou} = \frac{M_i}{M_i - 1}\hat{\alpha}_{i,Si,plug-in}. \tag{7}$$

is unbiased and asymptotically normally distributed. However, since M_i generally exceeds 1,000 in microbiome studies, the difference between the Zhang & 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 et al. (2017) investigate the performance of the zero-replace and low-rank approach to estimating the Simpson index. The extrapolation approach of Hsieh et al. (2016) also applies to the Simpson index. We compare our proposal, which we call `DivNet`, with these approaches in Sections 5 and 6.

2.1.3. α-diversity with covariates

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. Recently, Arbel et al. (2016) proposed a nonparametric Bayesian model that exploits structure in W as well as incorporating covariate information. However, the method is computationally expensive, and at present, an implementation only exists for $p = 1$. We compare our method to the method of Arbel et al. (2016) with respect to both estimation error and computation time in Section 5. We also note the recent method of Ren et al. (2017), which incorporates an error model into ordination methods, an alternative to diversity analysis in summarizing compositional 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 ecosystems. Legendre & 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 ecosystems have similar compositions, while large values indicate that the relative abundances differ between ecosystems, or that few taxa are shared by the ecosystems. 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 & Curtis 1957) is defined as

$$\hat{\beta}_{ij,BC,plug-in} = 1 - 2\frac{\sum_{q\in\mathcal{C}_i\cup\mathcal{C}_j}\min(W_{iq}, W_{jq})}{M_i + M_j}. \tag{8}$$

While we have not found any discussion of the target estimand in the literature, Eq. (8) suggests that

$$\beta_{ij,BC} = 1 - \sum_{q\in\mathcal{C}}\min(Z_{iq}, Z_{jq}) \tag{9}$$

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 et al. (2016) focused on estimating α-diversity, because their method estimates the latent composition matrix Z, we also compare our proposed method to the estimate

$$\hat{\beta}_{ij,BC,Arbel} = 1 - \sum_{q\in\mathcal{C}}\min\left(\hat{Z}_{iq}^{(Arbel)}, \hat{Z}_{jq}^{(Arbel)}\right), \tag{10}$$

where $\hat{Z}^{(Arbel)}$ is the latent composition matrix estimate based on the procedure of Arbel et al. (2016).

2.2.2. *Euclidean distance*

Finally, we mention the Euclidean distance between the relative abundance vectors,

$$\beta_{ij,ED} = \sqrt{\sum_{q\in\mathcal{C}}(Z_{iq} - Z_{jq})^2}, \tag{11}$$

whose plug-in estimate is

$$\hat{\beta}_{ij,ED} = \sqrt{\sum_{q\in\mathcal{C}_i\cup\mathcal{C}_j}\left(\frac{W_{iq}}{M_i} - \frac{W_{jq}}{M_j}\right)^2}. \tag{12}$$

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

$$\hat{\beta}_{ij,ED,Arbel} = \sqrt{\sum_{q\in\mathcal{C}}(\hat{Z}_{iq}^{(Arbel)} - \hat{Z}_{jq}^{(Arbel)})^2}. \tag{13}$$

3. Estimating diversity in networked composition data

Members of ecological communities interact, displaying repeatable patterns in many different environmental settings (Faust & 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 & Alm 2012) and SPIEC-EASI (Kurtz et al. 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. To our knowledge, this is the first method that explicitly accounts for co-occurrence patterns in diversity estimation.

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 (see also Mandal et al. (2015)).This models the counts W_{iq} as independent draws from a multinomial distribution,

$$p(W|Z) \propto \prod_{i=1}^{n} \prod_{q=1}^{Q} Z_{iq}^{W_{iq}}, \tag{14}$$

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

$$Y_{iq} = \phi(Z_{iq}) = \left\{ \log\left(\frac{Z_{iq}}{Z_{iD}}\right) \right\}_{q=1,\ldots,D-1,D+1,\ldots Q}. \tag{15}$$

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

$$Z_{iq} = \phi^{-1}(Y_{iq}) := \left\{ \begin{array}{ll} \dfrac{\exp(Y_{iq})}{\sum\limits_{q \neq D} \exp(Y_{iq}) + 1} & q \neq D \\ \dfrac{1}{\sum\limits_{q \neq D} \exp\{Y_{iq}\} + 1} & q = D. \end{array} \right\} \tag{16}$$

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

$$f(\mathbf{Y}_i|\mu, \Sigma) \propto |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2} (\mathbf{Y}_i - \mu_i)^T \Sigma^{-1} (\mathbf{Y}_i - \mu_i) \right\}. \tag{17}$$

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

3.2. Estimating diversity in the presence of a network

We propose using the log-ratio model to estimate α-diversity and β-diversity. Let $\hat{\beta}$ be an estimate of β under the log-ratio model. We discuss maximum likelihood estimators in detail in Section 4.1, and penalized maximum likelihood estimators in Section 4.2.

Suppose wish to estimate the α-diversity of an ecosystem with covariate vector $X_i \in \mathbb{R}^p$. Define

$$\hat{Y}_i = X_i^T\hat{\beta}, \tag{18}$$

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 $f : \mathbb{S}^{C-1} \to \mathbb{R}$:

$$\hat{\alpha}_i = f(\hat{Z}_i). \tag{19}$$

More explicitly,

$$\hat{\alpha}_{i,Sh,proposed} = -\sum_q \hat{Z}_{iq}\log\hat{Z}_{iq}, \tag{20}$$

$$\hat{\alpha}_{i,Si,proposed} = \sum_q (\hat{Z}_{iq})^2 \tag{21}$$

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

$$\hat{\beta}_{ij} = g(\hat{Z}_i, \hat{Z}_j), \tag{22}$$

such as

$$\hat{\beta}_{ij,BC,proposed} = 1 - \sum_q \min(\hat{Z}_{iq}, \hat{Z}_{jq}), \tag{23}$$

$$\hat{\beta}_{ij,ED,proposed} = \sqrt{\sum_q (\hat{Z}_{iq} - \hat{Z}_{jq})^2} \tag{24}$$

for the Bray-Curtis and Euclidean diversity indices. Note that if $\hat{\beta}$ is the maximum likelihood estimate of β, then by invariance, the proposed estimates are the maximum likelihood estimates 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 population from which the sample was drawn. This means that information is shared across all samples to obtain more precise and accurate estimates (see Section 5). In addition, samples i and j such that $||X_i - X_j||_\infty = 0$ will have $\hat{\alpha}_i = \hat{\alpha}_j$ and $\hat{\beta}_{ik} = \hat{\beta}_{jk}$ for any other sample k. In this way, biological replicates (samples where $||X_i - X_j||_\infty = 0$) have equal diversity index estimates, in contrast to plug-in estimates and the estimates of Chao & Shen (2003), Hsieh et al. (2016), and Cao et al. (2017). Furthermore, we can use the model to estimate the diversity of ecosystems for which ecosystem survey data is not available but for which covariate information exists. While these advantages are shared with the method of Arbel et al. (2016), our method is substantially faster (Figure 3), 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 $\eta = (\beta, \Sigma)$, we consider a maximum likelihood approach. If Y were known, our optimization problem would be to find

$$\hat{\eta} = \underset{\eta}{\operatorname{argmax}} \sum_{i=1}^{n} \left[\log Pr(\mathbf{W}_i|\mathbf{Y}_i) + \log f(\mathbf{Y}_i|\eta)\right], \tag{25}$$

where

$$\log Pr(\mathbf{W}_i|\mathbf{Y}_i) = \sum_{q \neq D} W_{iq} Y_{iq} - M_i \log\left(\sum_{q \neq D} \exp(Y_{im}) + 1\right) \tag{26}$$

and

$$\log f(\mathbf{Y}_i|\eta) = -\frac{1}{2}\log(|\Sigma|) - \frac{1}{2}(\mathbf{Y}_i - \mu_i)\Sigma^{-1}(\mathbf{Y}_i - \mu_i). \tag{27}$$

Alas, since Y is a latent random variable, we cannot directly optimize Eq. (25). Instead, we use the Expectation-Maximization algorithm (Dempster et al. 1977). The expected complete log-likelihood is

$$Q(\eta|\eta^{(t)}) = -\frac{n}{2}\log(|\Sigma|) - \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}_{Y|(W,\eta^{(t)})}[(\mathbf{Y}_i - \mu_i)^T\Sigma^{-1}(\mathbf{Y}_i - \mu_i)]. \tag{28}$$

To estimate this expectation numerically, we follow Xia et al. (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:

$$\mathbb{E}_{Y|(W,\eta^{(t)})}[(\mathbf{Y}_i - \mu_i)^T \Sigma^{-1} (\mathbf{Y}_i - \mu_i)] \approx \frac{1}{R} \sum_{r=1}^{R} (\mathbf{Y}_i^{(r)} - \mu_i^{(t)})^T (\Sigma^{(t)})^{\dagger} (\mathbf{Y}_i - \mu_i^{(t)}), \tag{29}$$

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 Eq. (28), 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:

$$\beta^{(t+1)} = \frac{1}{R} \sum_{r=1}^{R} \left[(X^T X)^{\dagger} X^T Y^{(r)}\right], \tag{30}$$

$$\mu_i^{(t+1)} = X_i^T \beta^{(t+1)}, \tag{31}$$

$$\Sigma^{(t+1)} = \frac{1}{nR} \sum_{r=1}^{R} \sum_{i=1}^{n} \left(\mathbf{Y}_i^{(r)} - \mu_i^{(t)}\right) \left(\mathbf{Y}_i^{(r)} - \mu_i^{(t)}\right)^T, \tag{32}$$

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 $t = 10$ is generally sufficient to achieve stable estimates. We run the Metropolis-Hastings 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 esti-

mates. These variance estimates can then be used in hypothesis testing (e.g., using the method of Willis et al. (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 $\hat{\beta}$ and $\hat{\Sigma}$ be the estimated values of β and Σ estimated by the algorithm described in Section 4.1.

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

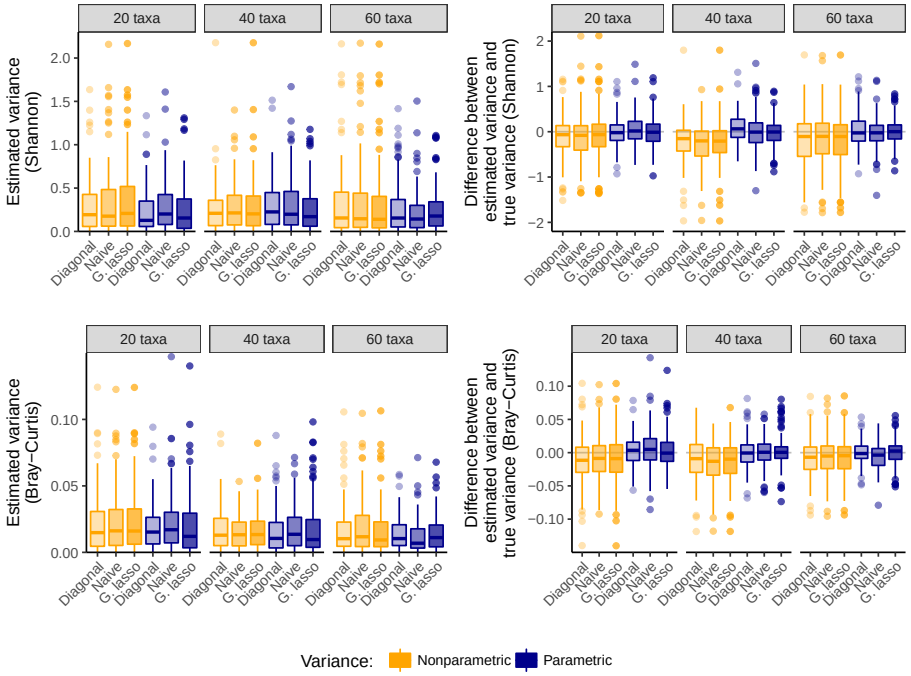
We also consider a nonparametric bootstrap approach to estimating the variance of our estimates. We uniformly at random select with replacement n_{sub} elements from $\{1, \ldots, n\}$ to obtain a set which we call $\mathcal{B}$. We then estimate $(\hat{\beta}^{(\mathcal{B})}, \hat{\Sigma}^{(\mathcal{B})})$ 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 $\{(\hat{\beta}^{(\mathcal{B})}, \hat{\Sigma}^{(\mathcal{B})})\}$ 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 non-parametric bootstrap estimate $\widehat{Var}(\hat{\alpha}_i) = \widehat{Var}_b(\hat{\alpha}_i^{(\mathcal{B}_b)})$ (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 samples obtained in microbiome surveys, and in this setting, $(\Sigma^{(t)})^\dagger$ may be a poor estimate of Σ^{-1} in Eq. (29), even for large t. We therefore consider replacing $(\Sigma^{(t)})^\dagger$ in Eq. (29) with a regularized estimate obtained from the graphical lasso (Friedman et al. 2008, Witten et al. 2011). Following the popular microbial network estimation software SPIEC-EASI (Kurtz et al. 2015), we use stability selection to select the regularization parameter (Liu et al. 2010, Kurtz et al. 2015). We also consider replacing $(\Sigma^{(t)})^\dagger$ with the maximum likelihood estimate 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 (2 approaches to estimating the variance for each of 3 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 $p = 1$, $n = 12$, and set $X = (\mathbf{1}_n^T, (\mathbf{0}_{2n/3}, \mathbf{1}_{1n/3})^T)$. Let $\mathcal{W}^Q$ be the columns of the count ma-

trix 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 $\beta^Q = (X^TX)^{-1}X^TY^Q$ and Σ^Q to be the covariance of the columns of $Y^Q - X\beta^Q$, and for each Q, we simulate data according to the log-ratio model with parameters β^Q, Σ^Q and $M_i = \sum_q W_{iq}$. Specifically, to simulate from the log-ratio model with parameters (β, Σ, X, M), we first simulate a matrix $Y \in \mathbb{R}^{n\times(Q-1)}$ with ith row $\mathbf{Y}_i \sim \mathcal{N}(X_i^T\beta, \Sigma)$, then calculate the matrix Z with ith row $\mathbf{Z}_i = \phi^{-1}(\mathbf{Y}_i)$ (see Eq. (15)), 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 $n_{sub} = 6$ for the nonparametric bootstrap approach.

FIG 1. *A comparison of candidate 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.*



We compare the estimated variance of the 6 methods in Figure 1 for a varying number of taxa Q. For brevity, only the variance of the Shannon index and Bray-Curtis index are shown. 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 understate 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 (β^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 3 approaches to covariance estimation show substantial advantage over the others. This 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

Having established estimators for diversity and variance, we now compare the performance of `DivNet` to estimates obtained from other methods. We simulate from the log-ratio model by specifying $\beta \in \mathbb{R}^{p\times Q}$, $X \in \mathbb{R}^{n\times p}$, $\Sigma \in \mathbb{R}^{Q\times Q}$, and $M \in \mathbb{R}^n$ and simulating W as described in Section 4.2.

Note that the true relative abundance vector for sample i is $\mathbf{Z}_i = \phi^{-1}(X_i^T\beta)$, and so the true diversity indices can be calculated for each choice of X and β. For each of the 4 diversity indices that we consider in this paper (Shannon, Simpson, Bray-Curtis, and Euclidean), we obtain an estimate under the multinomial model and using the proposed estimation procedure. The procedure of Arbel et al. (2016) can be applied when $p = 2$, and so we set $p = 2$ and choose $X = (\mathbf{1}_n^T, (\mathbf{0}_{n/2}, \mathbf{1}_{n/2})^T)$ for all simulations. The `R` package `iNEXT` (Hsieh et al. 2016) applies to estimating Shannon and Simpson α-diversity indices, but not to estimating β-diversity indices. Note that many of the Shannon diversity estimates are almost identical to the Multinomial MLE for large values of M (M is commonly 10^5 or greater in microbiome studies), including the estimates of Chao & Shen (2003) and Miller (1955), and for this reason we do not compare them here. For the same reason we also do not show the Simpson diversity estimate of Zhang & Zhou (2010). We use the `simulator` (Bien 2016) to manage the simulation study.

Throughout this section we evaluate α-diversity estimates using the mean square error (MSE) over all samples. The MSE of the kth simulation is

$$MSE_\alpha(\hat{D}^{(k)}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{D}_i^{(k)} - D_i)^2 \tag{33}$$

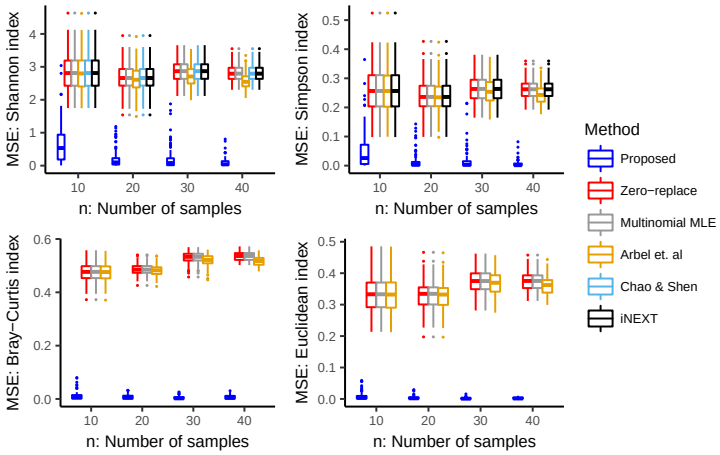
where i indexes the estimates for each of the n samples. We similarly evaluate the β-diversity estimates:

$$MSE_\beta(\hat{D}^{(k)}) = \frac{1}{n(n-1)/2}\sum_{i<j}(\hat{D}_{ij}^{(k)} - D_{ij})^2. \tag{34}$$

5.1. Estimation error decreases with sample size

We simulate the elements of $\beta_{pq} \overset{iid}{\sim} \mathcal{N}(0,1)$. To construct Σ, we construct a matrix $A \in \mathbb{R}^{(Q-1)\times(Q-1)}$ with elements drawn from a Uniform$(-1,1)$ distribution, and construct a diagonal matrix D with diagonal elements forming an arithmetic sequence of length Q beginning at σ_{max} and decreasing to a minimum of σ_{min}. We then set $\Sigma = A^T D A$. In this section we set $Q = 20$, $\sigma_{min} = 0.01$, $\sigma_{max} = 5$, and $M_i = 10^5$ for all i, and perform $K = 200$ simulations.

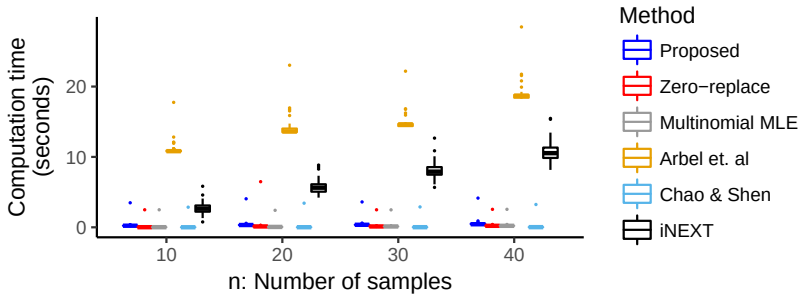
FIG 2. *A comparison of the error of different estimators for α- and β-diversity for communities where the taxa are networked. When the network is ignored by the estimation procedure (e.g., Chao & Shen (2003), Hsieh et al. (2016) and the widely used "plug-in" estimate (multinomial MLE)), the error in estimating diversity can be substantial. The proposed estimation procedure, which specifically accounts for networks, outperforms other estimates for any sample size n. The distribution of mean squared errors (MSEs) is shown for 200 simulated datasets. In this simulation, there are $M = 10^5$ microbes observed per sample, $p = 2$ predictors and $Q = 20$ taxa.*



The performance of the proposed method for estimating diversity when data is simulated under this model is illustrated in Figure 2. For all values of n and all diversity estimands, the 25%, 50%, and 75% quantiles of $\{MSE(\hat{D}^{(k)})\}_k$ are uniformly lower for our proposed method compared to all other methods. The improvement is especially pronounced for the β-diversity indices.

We find that the estimation error decreases as the sample size n increases for the proposed method and the method of Arbel et al. (2016), but not for

FIG 3. *A comparison of the computing time of different estimators of diversity indices. Our parallelized EM-MH algorithm for estimation under a network model is competitive with closed-form estimates, and is substantially faster to compute than the rarefaction-extrapolation approach of iNEXT (Hsieh et al. 2016) and the nonparametric Bayesian approach of Arbel et al. (2016). The computation time of the 200 datasets used to produce Figure 2 is shown.*



the Multinomial MLE and the iNEXT method (Figure 2). This is unsurprising, since neither the plug-in nor iNEXT estimates use information contained in the covariate matrix X in their estimates of diversity. Therefore, the additional information afforded by larger values of n is not leveraged by the plug-in nor iNEXT estimates, even when experimental replicates are available.

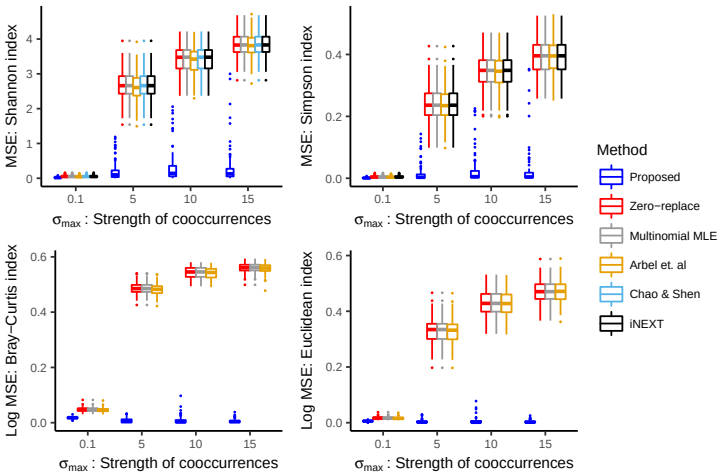
The results shown in Figure 2 are based on fitting our model with $t = 10$ EM steps and $r = 200$ MH draws per EM step. For these choices, we show computation time in Figure 3. Fitting our model with $t = 10$ EM steps and $r = 200$ is more computationally expensive than calculating the plug-in estimate, but less computationally expensive than fitting the model of Arbel et al. (2016) (with the default 10 chains) or using the package `iNEXT` (with default 40 knots and 50 bootstrap resamples). We note that our implementation leverages the `R` package `parallel` (R Core Team 2017) for parallelizing the MH algorithm employed at each E-step of the EM algorithm.

5.2. Estimation error is stable across networked communities

We now investigate the effect of varying the co-occurrence structure in the community on the estimation of diversity. Since larger values of the elements of Σ correspond to more strongly co-varying microbial abundances, by varying the elements of Σ, we can investigate the effect of the microbial concurrence network on diversity estimation. To vary the covariance structure in a systematic way, we vary σ_{max}, the largest eigenvalue of Σ. We generate β and X as in Section 5.1, set $n = 20$, $Q = 20$, $\sigma_{min} = 0.01$, $M_i = 10^5$ for all i, and perform $K = 100$ iterations for each choice of σ_{max}. The results are shown in Figure 4. We see that estimating the diversity in microbial communities with strong occurrence structures is more challenging than estimating diversity in communities with co-occurrence structure similar to that of a multinomial model. However, the proposed method has lower MSE than all other methods that were investigated.

Additionally, even when microbial abundances are simulated under a model with strong co-occurrence relationships, the proposed method can estimate the diversity with small MSE (Figure 4). In contrast, the estimation error increases as the co-occurrence relationships strengthen for all other methods. Co-occurrence relationships in microbial ecosystems are well documented (Faust & Raes 2012), indicating that a diversity estimation method tailored to networked ecosystems is of practical utility.

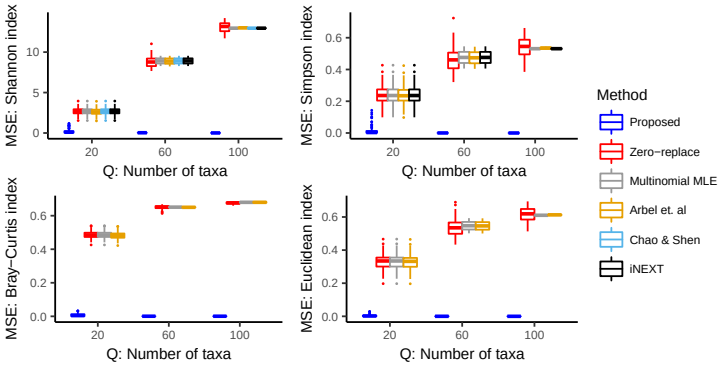
FIG 4. *Diversity estimates that incorporate network structure dominate estimators that do not incorporate network structure in the presence of a strong co-occurrence network. However, network-based estimates perform well even when there is a very weak network structure. As* $\sigma_{max} \to 0$*, the network model converges to the multinomial model. However, we see that the proposed network model performs equally as well or better than estimates based on the multinomial model for all choices of* σ_{max}*. This appears to be the case for estimating both* α*- and* β*-diversity.*



5.3. Estimation error is stable across large communities

Finally, since microbial communities often contain many taxa, we wish to confirm the performance of our estimator in large communities. In Figure 5, we see that the estimation error for the proposed method remains low even as the size of the community increases, while all other methods have increasing estimation error. In particular, we note that this is true even though the simulated communities are networked ($\sigma_{max} = 5$), and the number of taxa exceeds the number of samples ($n = 20$). We therefore conclude that the procedure is appropriate for analyzing the diversity of communities with many taxa, such as microbial communities.

FIG 5. *Diversity estimates that incorporate network structure dominate estimators that do not incorporate network structure over communities of any size. While most estimators have increasing error for larger communities, the proposed estimator's error does not. In the simulation, we set* $n = 20$ *and* $\sigma_{max} = 5$.



6. Data analysis: Seafloor microbial diversity

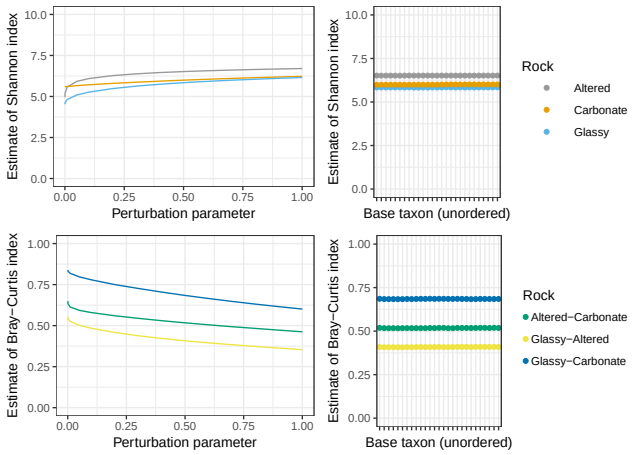
Because of its coarse nature as a community-level summary, diversity analyses are especially relevant to studies of novel ecological communities. Lee et al. (2015) collected and analyzed microbial communities living on seafloor rocks on the Dorado Outcrop, an area of exposed basalt on the East Pacific 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 lithified carbonate ("carbonate," $n = 1$), glassy, altered basalts ("glassy," $n = 4$), and highly altered basalts ("altered," $n = 8$). Analysis of the microbial communities on these rocks revealed 1490 distinct microbial taxa after filtering for low quality sequences (see Lee et al. (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 Q-th 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 66 amplicon sequence variants (ASVs) that were present in all samples, and so we uniformly at random select 10 ASVs from this collection of 66 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, i.e., let $U_j = \sum_{i=1}^{n} W_{ij}$, and call taxon d the k-th most abundant if $U_{(k)} = U_d$ for $U_{(k)}$ the k-th order statistic of the $\{U_j\}_j$. We also compare 10 randomly selected taxa. The estimated Shannon, Simpson, Jaccard, and Euclidean diversities are shown in Figure 6 (right panels), indicating that, in practice, the diversity estimates are almost invariant to the choice of base

taxon. Hereafter 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. As noted previously (Martín-Fernández et al. 2003, Cao et al. 2017), W_{ij} is commonly zero for microbiome data, because many taxa do not occur in every sample (46% 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 $p \in (0, 1)$ to create a new abundance table $W_{ij}^{(p)} = W_{ij} + p$, and the modeling the perturbed data $W^{(p)}$. In Figure 6, we observe sizeable changes in the diversity estimates when varying p close to zero (at most 24%, -274%, -36% and -53% changes in Shannon, Simpson, Bray-Curtis, and Euclidean estimates for $p = 0.001$ compared to $p = 0.5$), but smaller changes when p is increased from 0.5 to 1 (at most 5%, -30%, -15% and -15% changes for $p = 0.5$ to $p = 1$). We therefore follow Cao et al. (2017) and choose $p = 0.5$ as the perturbation parameter for the remainder of our analysis.

FIG 6. *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 is generally sparse, and 46% of the observed abundances of the Lee et al. (2015) dataset are zero. For this reason, it is common to add a perturbation offset p to the observed abundance table before fitting the log-ratio model. Here we see that the estimated diversity does depend on the choice of p.*



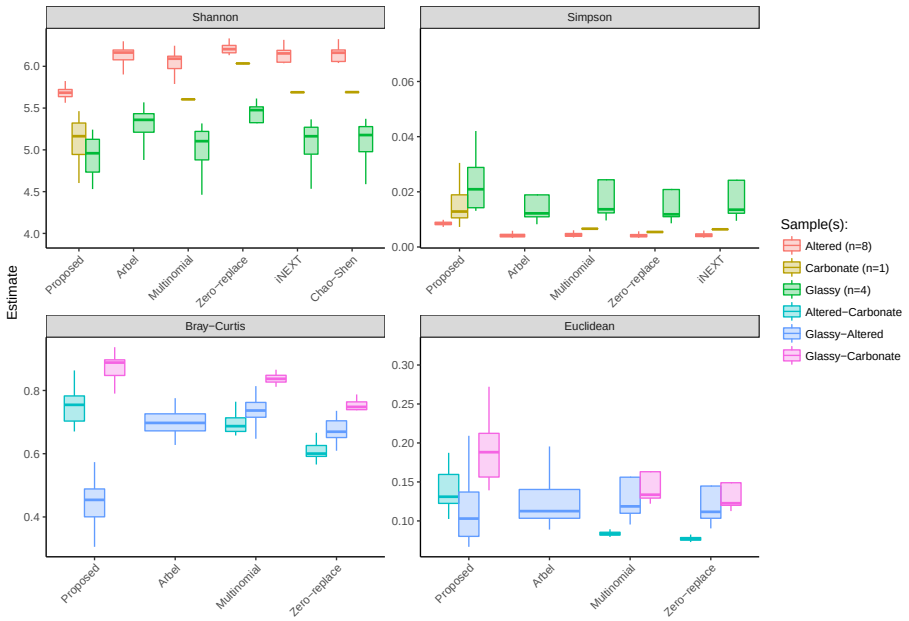
Throughout this paper we have argued that the multinomial model is misspecified for microbiome data. To investigate this claim for the dataset of Lee et al. (2015), we fit the log-ratio model and calculate the eigenvalues of of $\hat{\Sigma}$. A test that that the largest eigenvalue of Σ is zero is rejected with $p < 10^{-5}$ ($\hat{\ell}_1 = 4610.44$, $\mu_{np} = 1779$, $\sigma_{np} = 28.33$) (Johnstone 2001, El Karoui 2003). This is strong evidence that a networked model is appropriate for this dataset.

Finally, we compare the estimates obtained from our method to the estimates obtained from other methods. Interval estimates are shown in Figure 7. While

most methods produce similar estimates, we note a number of advantages of our proposal. Firstly, our method handles multiple covariates, which the method of Arbel et al. (2016) does not. Secondly, any diversity index that is a function of relative abundance can be estimated using our method, unlike the methods of Hsieh et al. (2016) and Chao & Shen (2003). Thirdly, our interval estimates are more symmetric around the median of the bootstrapped estimates compared to other estimates, indicating the greater stability of `DivNet` compared to other methods.

The final advantage that we note is that for a sample condition which was only observed once (carbonate), an interval estimate is computable and of nonzero length. Our method is the only method that achieves this. The method of Arbel et al. (2016) cannot handle multiple covariates, and the remaining estimators can only produce point estimates based on a single sample. It is also worth noting that the interval for the sample observed once is wider than the interval for samples observed more than once (altered and glassy basalts), which is consistent with the amount of information available about this sample condition.

FIG 7. *Lee et al. (2015) collected and analyzed microbial communities living on 3 types of seafloor basalts on the Dorado Outcrop. Here we compare a variety of estimators for 4 diversity indices (25% and 75% quantiles are shown). Our method works with multiple covariates, produces approximately symmetric interval estimates, and is the only method which can produce confidence intervals for the carbonate basalts.*



7. Discussion

Despite substantial evidence that strong co-occurrence networks exist in ecological 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. We have shown that `DivNet` is accurate, fast, performs well with a large number of taxa, and incorporates replicate and covariate information. It also permits extrapolation to experimental conditions that were not observed. It is available as a `R` package via `github.com/adw96/DivNet`.

By leveraging information from multiple samples, `DivNet` can estimate the relative abundance of a taxon in an ecosystem where it was not observed. However, a limitation of `DivNet` is 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 et al. (2016) and Cao et al. (2016), while the estimators of Hsieh et al. (2016) and Chao & Shen (2003) adjust for missing taxa (but are only applicable to α-diversity). However, the latter 2 estimators cannot handle covariates nor repeated samples, which we believe significantly contributes to the strong performance of our method. We note that 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.

We suggest 3 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, such as sparsity, on the relative abundance parameter β, whose dimension is large when there are a large number of taxa. Finally, since diversity indices that simultaneously incorporate relative abundance and phylogenetic information are commonly used by ecologists, extending the method to incorporate phylogeny is a challenging open problem.

All code to reproduce the simulations and data analysis, along with tutorials for using our package, are available at `github.com/adw96/DivNet`.

Acknowledgements

The authors are 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.

References

Aitchison, J. (1982), 'The statistical analysis of compositional data', J Roy Stat Soc B Met **44**.

Aitchison, J. (1986), 'The statistical analysis of compositional data'.

Arbel, J., Mengersen, K. & 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**(3), 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**(3), 333–336.

Bien, J. (2016), 'The Simulator: An Engine to Streamline Simulations', arXiv preprint arXiv:1607.00021 .

Billheimer, D., Guttorp, P. & Fagan, W. F. (2001), 'Statistical interpretation of species composition', Journal of the American Statistical Association **96**(456), 1205–1214.

Bray, J. R. & Curtis, J. T. (1957), 'An ordination of the upland forest communities of southern Wisconsin', Ecological Monographs **27**(4), 325–349.

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

Cao, Y., Lin, W. & Li, H. (2016), 'Large Covariance Estimation for Compositional Data via Composition-Adjusted Thresholding', arXiv preprint arXiv:1601.04397 .

Cao, Y., Zhang, A. & Li, H. (2017), 'Microbial Composition Estimation from Sparse Count Data', arXiv preprint arXiv:1609.03045 .

Chao, A. & Shen, T.-J. (2003), 'Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample', Environmental and Ecological Statistics **10**(4), 429–443.

Chao, A., Wang, Y. T. & 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**(11), 1091–1100.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', Journal of the royal statistical society. Series B (methodological) pp. 1–38.

El Karoui, N. (2003), 'On the largest eigenvalue of Wishart matrices with identity covariance when n, p and p/n go to infinity ', arXiv preprint math/0309355 .

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

Faust, K. & Raes, J. (2012), 'Microbial interactions: from networks to models', Nature Reviews Microbiology **10**(8), 538–550.

Friedman, J. & Alm, E. J. (2012), 'Inferring correlation networks from genomic survey data', PLoS Computational Biology **8**(9), e1002687.

Friedman, J., Hastie, T. & Tibshirani, R. (2008), 'Sparse inverse covariance estimation with the graphical lasso', Biostatistics **9**(3), 432–441.

Gibson, K. L., Wu, Y.-C., Barnett, Y., Duggan, O., Vaughan, R., Kondeatis, E., Nilsson, B.-O., Wikby, A., Kipling, D. & Dunn-Walters, D. K. (2009), 'B-cell diversity decreases in old age and is correlated with poor health status', Aging cell **8**(1), 18–25.

Hsieh, T. C., Ma, K. H. & Chao, A. (2016), 'iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers)', Methods in Ecology and Evolution **7**(12), 1451–1456.

Johnstone, I. M. (2001), 'On the distribution of the largest eigenvalue in principal components analysis', Annals of Statistics pp. 295–327.

Kaplinsky, J. & Arnaout, R. (2016), 'Robust estimates of overall immune-repertoire diversity from high-throughput measurements on samples', Nature communications **7**, 11881.

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J. & Bonneau, R. A. (2015), 'Sparse and Compositionally Robust Inference of Microbial Ecological Networks', PLoS Computational Biology **11**(5), 1–25.

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

Lee, M. D., Walworth, N. G., Sylvan, J. B., Edwards, K. J. & 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**.

Legendre, P. & Legendre, L. F. (2012), Numerical ecology, Vol. 24, Elsevier.

Liu, H., Roeder, K. & Wasserman, L. (2010), Stability approach to regularization selection (stars) for high dimensional graphical models, in 'Advances in neural information processing systems', pp. 1432–1440.

Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. (2012), 'Diversity, stability and resilience of the human gut microbiota', Nature **489**(7415), 220.

Lozupone, C. & Knight, R. (2005), 'UniFrac: a New Phylogenetic Method for Comparing Microbial Communities', Applied and Environmental Microbiology **71**(12), 8228–8235.

Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R. & Peddada, S. D. (2015), 'Analysis of composition of microbiomes: a novel method for studying microbial composition', Microbial ecology in health and disease **26**(1), 27663.

Martín-Fernández, J. A., Barceló-Vidal, C. & Pawlowsky-Glahn, V. (2003), 'Dealing with zeros and missing values in compositional data sets using nonparametric imputation', Mathematical Geology **35**(3), 253–278.

McCoy, C. O. & 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**(95), 100.

Oakley, B. B., Fiedler, T. L., Marrazzo, J. M. & Fredricks, D. N. (2008), 'Diversity of human vaginal bacterial communities and associations with clinically defined bacterial vaginosis', Applied and Environmental Microbiology **74**(15), 4898–4909.

R Core Team (2017), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria.

Ren, B., Bacallado, S., Favaro, S., Holmes, S. & Trippa, L. (2017), 'Bayesian

nonparametric ordination for the analysis of microbial communities', Journal of the American Statistical Association **112**(520), 1430–1442.

Shannon, C. E. (1948), 'A Mathematical Theory of Communication', Bell System Technical Journal **27**(3), 379–423.

Simpson, E. H. (1949), 'Measurement of diversity.', Nature .

Vu, V. Q., Yu, B. & Kass, R. E. (2007), 'Coverage-adjusted entropy estimation', Statistics in Medicine **26**(21), 4039–4060.

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E. R. & Knight, R. (2017), 'Normalization and microbial differential abundance strategies depend upon data characteristics', Microbiome **5**(1), 27.

Willis, A. (2017), 'Rarefaction, alpha diversity, and statistics', bioRxiv .

Willis, A., Bunge, J. & Whitman, T. (2016), 'Improved detection of changes in species richness in high-diversity microbial communities', Journal of the Royal Statistical Society: Series C **66**(5), 963–977.

Witten, D. M., Friedman, J. H. & Simon, N. (2011), 'New Insights and Faster Computations for the Graphical Lasso', Journal of Computational and Graphical Statistics **20**(4), 892–900.

Xia, F., Chen, J., Fung, W. K. & Li, H. (2013), 'A logistic normal multinomial regression model for microbiome compositional data analysis', Biometrics **69**(4), 1053–1063.

Zahl, S. (1977), 'Jackknifing an index of diversity', Ecology **58**(4), 907–913.

Zhang, Z. & Zhou, J. (2010), 'Re-parameterization of multinomial distributions and diversity indices', Journal of Statistical Planning and Inference **140**(7), 1731–1738.