The power and pitfalls of Dirichlet-multinomial mixture models for ecological count data

John D. O'Brien[1] and Nicholas Record[2]

[1]Bowdoin College, Department of Mathematics, Brunswick, Maine 04011 USA
[2]Bigelow Laboratory for Ocean Sciences, East Boothbay Harbor, Maine 04544 USA

March 23, 2016

Abstract

The Dirichlet-multinomial mixture model (DMM) and its extensions provide powerful new tools for understanding the ecological dynamics underlying taxa abundance. However, their capacity to effectiveness in capturing features of empirical data is not well understood. In this work, we use posterior predictive distributions (PPDs) to explore the performance of an infinite dimensional version of the DMM in three case studies, including two amplicon metagenomic time series. We avoid concentrating on fluctuations within individual taxa and instead focus on consortial-level dynamics with straight-forward methods for visualizing the output of the model that emphasizes this perspective. In each study, the DMM appears to perform well in organizing the data as a framwork for biological interpretation. Using the PPDs, we also observe several exceptions where the data appear to significantly depart from the model in ways that give useful ecological insight. We summarize the conclusions as a set of considerations for field researchers: problems with samples and taxa; relevant scales of ecological fluctuation; additional niches as outgroups; and possible violations of niche neutrality.

Introduction

Metagenomic and amplicon sequencing techniques provide an unpredecented vantage to investigate the dynamics of microbial ecologies, allowing researchers to simultaneously observe counts of thousands of species and will undoubtedly mold advances in human health, environmental science, and engineering [53, 61, 65]. The ability to observe whole ecosystems *in situ* has elevated microbiomes to a central position within ecology [66, 29, 1, 11, 15], These massive data sets present researchers with a new challenge: how to understand what they communicate about microbial ecological dynamics and, in turn, how this might inform ecology more broadly. This new project requires a dramatic shift in perspective, away from dynamics of small numbers of species to the interactions of thousands of them.

While metagenomic data adds a new urgency, taxa abundance data or ecological count data – collections where researchers estimate the number of observed individuals across a set of taxa within an ecosystem – is one the oldest forms of ecological measurement. Understanding the underlying phenomena governing how taxa abundance varies in time, space or along other environmental gradients is a foundational challenge within ecology [9, 18]. A wide wide number of theoretical perspectives attempt to understand these variations, such as Lotka-Volterra and its descendants to models based on maximum entropy, food webs, trait-based dynamics, metacommunities, trophic cascades, mass and energy balance, and entropy production, among others [71, 37, 26, 3, 31, 32, 34, 8, 68, 49, 41, 2, 28, 47], While many have compelling empirical or theoretical features, it is hard to argue that any have yet risen to provide a unified theoretical/empirical framework after the fashion of the coalescent model in population genetics [56, 67].

One of the most promising avenues to meet this challenge is the unified theory of neutral biodiversity (UNTB), pioneered by Hubbell, that builds on the concept of neutral niches – collections of species that are ecologically equivalent – within a metacommunity structure [26, 22]. While frequently controversial, the UNTB has provided a versatile framework for understanding the structure of species abundance, particularly for species richness, dispersal, and island biogeography [4, 16, 57]. The crucial insight for the analysis of ecological count data is that, under suitable conditions, the Dirichlet-multinomial (DM) distribution is the sampling distribution of a neutral niche [22]. If the count data for each sample are assumed to arise from a latent niche modeled by a DM distribution, the full metacommunity structure is given by a DM mixture model (DMM). [25] gave the first practical inference for the DMM based on a variational Bayesian approach. Recent work by these authors shows how this framework can be broaded to an infinite-dimensional framework that can test for metacommunity neutrality [22]. An independent extension uses a similar model for supervised feature detection [59].

The DMM is only one of many approaches to make sense of ecological count data in microbiome contexts. Unifrac, a phylogeny-based probability metric and other distance based methods dominanted early approaches and are still widely used [36, 13]. More recently, regression models based on single component DM and logistic normal distributions have given researchers a framework to understand how to associate significant shifts at the taxa level with environmental covariates while effectively accounting for structure of the overall count data [14, 70]. An entirely distinct avenue has explored network-based analysis of taxa correlations with increasing levels of statistical rigor [35, 7].

While appreciating the merits of these other approaches, here we concentrate just on how well the DMM captures the features of empircal data. For field researchers, the statistical truism that ‘all models are wrong, but some are useful,’ is naturally followed by the question ‘how useful are they?’ or, more pointedly, ‘how wrong are they?’ Our aim in this work is to give some answers to these questions for the DMM. Presented with a taxa abundance dataset, the DMM will always provide a means of organizing this data into distinct components; how well this organization can be relied upon is necessary consideration for empirical ecologists. While the work of [22] points to how the DMM may be re-embedded for general hypothesis testing, the integrated computational frameworks for statistical exploration common in other biological disciplines appear still far off [17, 55]. We emphasize that researchers can hold reasonable doubt about the verity of the UNTB, and still benefit from the DMM as productive means of organizing and interpreting ecological count data [52, 38]. This requires researchers to exercise due caution about the empirical features that may have be neglected or distorted by this analysis, such as ecologies dominated by a single species or bias caused by a small number of samples.

To measure how well the model captures features of the data, we use posterior predictive distributions (PPD) to compare between the empirically observed data and that data would be expected if the DMM were the correct underlying model [19, 39]. PPDs are an important tool in Bayesian analysis, and provide a crucial means of assessing how adequately complex models are able to capture features of their underlying data sets. For a model specified by parameters Θ and data $\mathcal{D}$, the posterior distrubution is the conditional distribution $\mathbb{P}(\Theta|\mathcal{D})$. The posterior predictive distribution for a new set of data $\widetilde{\mathcal{D}}$ is generated by integrating the model likelihood of $\widetilde{\mathcal{D}}$ over the posterior distribution:

$$\mathbb{P}(\widetilde{\mathcal{D}}|\mathcal{D}) = \int_{\Theta} \mathbb{P}(\widetilde{\mathcal{D}}|\Theta) \cdot \mathbb{P}(\Theta|\mathcal{D}) \cdot d\Theta.$$

For complex models where this integral cannot be done analytically this distribution can be approximated by simulating a large number of datasets $\widetilde{\mathcal{D}}_1, \cdots, \widetilde{\mathcal{D}}_N$ from $\mathbb{P}(\Theta|\mathcal{D})$. We then compare the observed data to these simulated sets through a test statistic T. Absent sufficient statistics, the appropriate choice of T is critical to reveal the empirical attributes of interest. These comparison are often summarized in terms of a posterior predictive p-value (PPP), the fraction of simulated test statistics more extreme than the empirically observed value:

$$\mathbb{P}\Big(T(\widetilde{\mathcal{D}}_i) > T(\mathcal{D})\Big) \approx \frac{1}{N}\sum_{i=1} \mathbf{1}_{[T(\widetilde{\mathcal{D}}_i) > T(\mathcal{D})]},$$

where $\mathbf{1}_{[.]}$ is an indicator function. Here, we choose a set of test statistics intended to capture common ecological concerns: within-sample taxa mean; across-sample mean for taxa; across-sample standard deviation for taxa; number of absent species within samples; and pairwise correlations among taxa.

Throughout the manuscript, we will refer to the components of the DMM as ecostates. [25] use the term 'ecotypes' for these consortia, which is easily confused with its more common usage for a geographically-adapted subpopulation within a species. These components have also been labeled 'metacommunities,' though that term could also stand in for the collection of components [43]. Though 'ecostates' already has a referent within ecology (a leaf without midvein), it is sufficiently distinct from the DMM and other closely-related topics that the meaning is clear from context. We believe this phrasing also emphasizes a statistically pragmatic perspective about what these components represent: consistently inferred taxa consortia from a set of ecological samples, rather than niches or metacommunities per se.

In this investigation, we first outline a new inference scheme the generalizes the DMM to make use of a Dirichlet process mixture model (DPM). In it's original form, the DMM provides a single estimate, K, of the number of ecostates that best explain a set of count data. However, in considering PPDs, a reasonable concern is that some model discrepancy arises purely from uncertainty in the choice of K rather than within the model itself. The DPM generalization of the DMM allows integration over all possible values of K and so directly accounts for this form of uncertainty. When necessary, we will refer to this generalization as the DPM-DMM. We provide the scripts for this model under a Creative Commons License, written in the open-source R computing environment, as part of the online supplementary material.

We then apply the DMM to three diverse time series data sets from across ecology: well-known microbial abundance data collected from two individuals sampled over nearly 450 days at four bodily locations; microbial abudance count data collected from the English channel approximately every two weeks over the year 2009-2010; and copepod abundance count data from continuous plankton recorders (CPR) collected during transects across the Gulf of Maine over the last fifty years. For each of these, we use simple visualizations to emphasize how the DMM reveals many previously identified features of the data, as well point to some novel insights. The PPD analysis measures the overall agreement within the datasets, and we highlight problematic features and possible explanations within each. We conclude with a discussion including guidelines for field researchers using these methods for analysis.

Data and Methods

Data

We use three publicly available environmental time series as data sets. Each series was filtered for quality of metadata and the total number of counts per taxa. For the copepod time series, we filtered the data into three different sets based on taxa abundance. The other two sets a single set was used that screened out low abundance taxa. The scripts filtering the original data are available in the online supplement to the paper.

Human-associated bacterial abundance time series

To our knowledge, the most intensive amplicon-based, publicly available time series comes from 396 time points collected from four locations on two human individuals, frequently known by the title of the paper that described the data: the 'Moving Pictures of the Human Microbiome (MPHM) [10].' This data was collected from the right and left palms, feces and saliva of a male and female subject over a total of 445 days from October 2008 until January 2010, for a total of 1967 samples. Illumina GAII sequencing targeted at the 16S rRNA V4 region was applied to each sample and the resulting reads mapped to the Greengenes reference set, with the full protocol specified in [10]. We downloaded the processed data and associated sample metadata from the Earth Microbiome Database on July 15, 2015. The data were filtered by species abundance for all taxa with more than 3100 counts across all samples, yielding 742 taxa in total.

English channel bacterial abundance times series

The L4 station is a sampling location for the Western Observatory of the English Channel and the site of one of the first longitudinal metagenomic marine sequencing projects [21, 20, 11]. This collection builds on an extensive record of ecological and environmental sampling at this location dating back to 1903, with continuous plankton recording since 1998[]. This time series is the publicly available portion of a larger six-year series, and contains 68 samples gathered approximately 2 weeks apart from April 2009-April 2010. The large majority of samples were collected in pairs. Amplicon 16S rRNA targeted at the V4 hypervariable region was applied to each sample, as well as metatranscriptomic sequencing, although we do not consider that data here. We downloaded the raw data set and metadata including sampling times from the MG-RAST database on March 15, 2013 [40].

We aligned the raw read data and screened for homopolymer artifacts using functions in the `mothur` software[58]. Reads were alligned to the SILVA reference alignment of 10,242 prokaryotic species (Release 102) [50]. These were translated into species counts At this stage, samples contained highly variable numbers of counts, from 2 to 267,529, with most samples near the median value of 33,041. We filtered this data into a single set for analysis, removing all taxa with fewer than 3 counts, leading to 994 taxa in total.

Gulf of Maine time series copepod abundance data

Zooplankton are a crucial link in the ocean food web between the fish populations and microbial-scale tropic levels. For the past 70 years, continuous plankton recorders (CPR) monitored their abundance and diversity, most frequently by being added opportunistically to ships transiting a region of interest. CPRs provide count data for identifiable copepod species at a depth of 5-10 meters below the ocean's surface and have been used widely due to their relative absence of expense and ease of deployment. Here, we consider the copepod data collected via CPRs from 1964 to 2010 during transits across the Gulf of Maine (GoM), running from near Boston, Massachusetts to Yarmouth, Nova Scotia, available at the COPEPOD database [46]. While CPR data is among the most extensive marine ecological measurements, comparisons with more metagenomic barcoding methodologies suggest that their total counts may be somewhat biased [33, 24].

The complete data contain 4799 samples, each enumerating the observed counts for a single CPR for 51 copepod species or genera. As counts ascertained by morphology, taxa assignments are not exclusive but not overlapping, so that each count is resolved to the lowest taxonomic level possible. Each sample posssesses metadata for the time of collection and longitude and latitude, as well as a record of phytoplankton color index, a proxy measurement for the phytoplankton levels at the time of sampling. There is wide variablility in the number samples for each year, from ten in 1977 to 145 in 1989, with most measurements occurring between 1980 and 2000 (2456/4799). Each sample is located along an approximate two dimensional transect across the GoM. We filtered the samples to be located within 150 kilometers of the central line of this transect, excluding 93 samples. An additional six samples were excluded due to ambiguous metadata.

Methods

Model

We employ an infinite dimensional generalizations of the multinomial-Dirichlet mixture model of Holmes *et al.* [25]. This allows us to infer a posterior distribution that integrates over values of K, the number of components, removing this as a factor the analysis. As in Holmes *et al.* we assume that each sample's count data arises from a DM distribution. This distribution allows for additional dispersion relative to a strict multinomial distribution. The model assumes that there are an unknown number of DM components (ecostates) underlying the data and that each sample comes from one of these components. Presuming the samples are otherwise exchangeable, a latent variable c_i augments each sample to assign it one of the ecostates. Supposing the DM paramers for an ecostate k are given by $A_k = (\alpha_{1k}, \cdots, \alpha_{Mk})$, then, conditional upon $c_i = k$, the likelihood for a sample is given by

$$\begin{aligned} \mathbb{P}(\mathcal{D}_i | c_i, A_k) &= \frac{\Gamma(\overline{A}_k)}{\Gamma(\overline{\mathcal{D}}_i + \overline{A}_k)} \prod_{r=1}^{M} \left(\frac{d_{ir} + \alpha_{kr}}{\alpha_{kr}} \right) \\ &= \mathrm{DM}(A_k) \end{aligned} \quad (1)$$

where Γ is the gamma function, $\overline{A}_k = \sum_{r=1}^{M} \alpha_{rk}$ and $\overline{\mathcal{D}}_i = \sum_{r=1}^{M} d_{ir}$. As the samples are assumed to be exchangeable, the full likelihood is then the product over all $\mathbb{P}(\mathcal{D}_i|c_i, A_k)$:

$$\mathbb{P}(\mathcal{D}|\mathbf{A}, \mathbf{c}) = \prod_{i=1}^{N} \mathrm{DM}(A_{c_i})$$

where $\mathbf{c} = (c_1, \cdots, c_N)$. Like [25], we adopt a Bayesian perspective on the problem. However, to remove K from consideration, we use a Dirichlet process mixture model (DPM), a nonparametric approach to specify the prior distribution on the parameters [42, 64]. Following [44], the DPM can be formulated in this context as:

$$\begin{aligned} \mathcal{D}_i|c_i = k &\sim \mathrm{DM}(A_k) \\ A_k|G &\sim G \\ G &\sim \mathrm{DP}(G_0, \eta) \end{aligned}$$

where $\sim$ denotes'distributed as', G_0 is the base measure, $\eta > 0$ is a concentration parameter, and DP is a Dirichlet process. The base measure here is the Cartesian product of two independent distributions, with the first component an exponential distribution with mean one and the second a uniform Dirichlet distribution of length M.

Inference

We use a Markov chain Monte Carlo (MCMC) methodology to approximate the posterior distribution of model, following the methods described in [44, 62]. We use two sets of Gibbs updates, one for the DPM parameters and one for the DM parameters. The DPM parameters (the latent variables) are drawn using a a collapsed Gibbs sampler. At each iteration, the collapsed Gibbs update successively moves the sample's latent variables to possibly new states, as specificed in Neal's Algorithm 8 [44]. For each sample i, we let l be the number of distinct culture labels c_s for $s \neq i$ and $h = l + m$ where m is a parameter that allows for the assignment of a sample to a number of new components. We fix $m = 3$. If $c_i = c_s$ for some $i = s$, then we sample values for that component. If $i \neq s$ for any s, then we set $c_i = l + 1$ and draw m new components from the base measure. Finally, we draw a new value for c_i according to:

$$\mathbb{P}(c_i = c|c_{-i}, A_1, \cdots, A_h) \propto \begin{cases} \frac{n_{-i,c}}{N-1+\eta} \cdot \mathrm{DM}(A_c) & \text{for } 1 \leq c \leq l, \\ \frac{\eta/m}{N-1+\eta} \cdot \mathrm{DM}(A_c), & \text{for } l < c \leq h, \end{cases} \tag{2}$$

where c_{-i} denotes all values of c except c_i and $n_{-i,c}$ is the number of values equal to c for $i \neq s$. At each iteration, the set of A_c's is renumbered so all components have at least one associated sample.

The Gibbs steps to update the DM parameters come from recent work on how to generate efficient sampling from this distribution [62]. The method relies on a data augmentation that separates the DM distribution into a Dirichlet distribution and a log-concave single parameter distribution. Since the Dirichlet is conjugage to the prior, it can be directly sampled. The last parameter, the dispersion parameters for the DM distribution, is sampled using a Griddy Gibbs sampler with 1000 draws at each iteration [54].

For each dataset we ran three MCMC chains for 20,000 iterations. In each MCMC run, we applied several diagnostics to ensure convergence, including autocorrelation plots, Geweke's diagnostic test, and estimating the effective sample size. We take a 10% of the chain as burn-in.

As an example, for one run of the full copepod data set, we realized minimal autocorrelation by thinning by 10. The Geweke statistic on the thinned chain was 2.047, consistent with reasonable convergence. The effective sample size was 423.78.

Posterior predictive distribution

To simulate the PPD, we used simulation routines in the R computing environment developed for the Human Microbiome Project [66]. For each MCMC run, we simulated data in the following fashion. We thinned the Markov chain by 10 and trimmed it with a burn in of 2,000 iterations. For each posterior sample and each data sample, we generated counts according to the corresponding DM distribution with the total number of counts equal to that observed in the data sample. We repeated this procedure 1,000 times, generating 50,000 simulated data sets per sample. We then calculated the PPP summaries for each taxa within each sample, the overal mean, overall standard deviation, number of zeros within a sample, and the pairwise correlations for the 15 most abundant and least abundant species using these simulated data sets. The PPP for each taxa within each sample was calculated by observing the fraction of simulated samples with more counts than the observed value. The mean, standard deviation, and pairwise correlation calculation were taken over all samples, providing metrics for each species. The zero calculation is taken over all species, provides a metric for each sample. All PPPs were then transformed by $\tilde{p} = |0.5 - p|/0.5$ so that values of either type of extremity yielded similar values.

Visualizing the posterior distribution

Visualizing complex data is a crucial means to harness understanding from scientific inquiries. The DPM-DMM effectively clusters the samples into a finite but variable number of components that reflect similarities in the count distributions across the samples. Most presentations using the DMM show taxa counts organized by these clusters to emphasize the specific shifts apparent across components [73, 12]. This is a natural approach particularly in a case-control study, where the researchers often seek to find taxa that separate treatments. However, in the time series here, we prefer a presentation to highlight the ecostates' relationship to each other and to time. This visualization is a simple organization of the posterior samples in a vein similar to the results of chromosome painting, so we term it 'ecological painting' [23].

An ecological painting is simply a plot that colors samples according to their ecostate and orders the samples in time and space. In principle, other covariates in principle may be used but these are natural first points of utilization. In the case of a finite mixture model, the number of components is fixed and so, up to label switching, the colors are assigned proportional to the fraction of posterior samples in that state. In the case, we need to identify a homology of the ecostates across the posterior samples.

Unfortunately, there is no intrinsic aspect of the ecostates that render them identifiable across iterations. We deal with this problem by rendering the posterior into a finite mixture in the following way. To this, we first generate a coherence plot of the MCMC chain, showing the frequency of how often different samples fall within the same component. We also histogram the distribution of the number of components across the chain. From these we then determine a minimum number of components, K', that reasonably capture the posterior sample. This value may not be the smallest number that the chain takes, and samples with components fewer that K' are excluded from the visualization. In our empirical examples, these iterations formed at most 2% of the chain. With

K' in hand, for each iteration we find the largest K' components, and then enumerate all other components as $K' + 1$. This effectively reduces the chain to the finite case and so we apply the scheme propopsed by Stephens to minimize the Kullback-Leibler divergence between components across iterations [63].

Researchers may also be interested in the how the ecostates relate to each other. To summarize this information, we present the expected frequency for the twenty most frequent taxa for each component, along with a hierarchical clustering constructed on the matrix of Canberra pairwise distance among the model components. We note that the hierarchical clustering can be sensitive to the metric used, although we find that the Canberra distance provides a reasonable measure to capture the relationships among expected frequencies.

Results

We present the results for each data set, beginning with the biological interpretation of the ecological painting, followed by the indications the PPD gives about the model fit, and the correlation in ecostates between different level of taxa filtering. For each data set, we produce two figures – the painting and a summary of the PPD – but only show those with novel results. For the remaining figures, we refer to the Supplementary Information.

Human-associated habitat bacterial time series

Figure 1 shows the ecological painting of the DPM-DMM applied to the most heavily filtered set of the MPHM data. This presentation distills a number of key results from the original paper, rendering complex data amenable to visual analysis [10]. Most salient is that the three niches (hand, feces, and tongue/saliva) are clearly separated by their inferred ecostates. Interestingly, the states assigned to tongue/saliva and hand habitats are shared across the two individuals, while, the feces habitat marks strong separation between them, consistent with the enterotypes hypothesis [69, 5]. For each habitat within each individual, the inferred ecostate exhibits strong temporal consistency. In particular, the female subject's fecal samples exhibit near perfect consistency in ecostate across all time points. The ecostate painting shows the two hand pairs for both individuals exhibit strong correlation in their ecostates ($\rho = 0.72$, for maximum likelihood of the male subject's samples), as observed in the original report.

The DPM-DMM model also supports a more refined analysis than a species-by-species comparison. The strong temporal consistency in ecostate assignment across all habitats is modulated by regular oscillation within some of them. Within both individuals, the tongue/saliva habitat oscillates between two ecostates, with the duration of the oscillations lasting approximately 1 month for the male subject, and less than a week for the female subject. The painting also identifies a number of unusual samples, most obviously those colored green in the male hand and fecal samples. These may be mislabeled samples, or show unexpected associations in human microbiomes (between saliva and hand, for instance).

Figure 2 summarizes the posterior predictive analysis for the data set. This indicates that the model effectively captures the mean and standard deviation in species counts and the frequency of zero counts in samples. The individual p-vales for each species across each sample (upper left panel) appear problematic but largely recapitulate the habitat structure, with reasonable p-values where particular species are common. The model struggles to capture correlations across species, both at high and low levels of association. Due to habitat structuring, the MPHM shows strong pairwise structure in correlation, with each pair of species showing high positive correlation ($\rho > 0.5$), low correlation ($-0.5 < \rho < 0.5$), or high negative correlation ($\rho < -0.5$). In nearly all cases, the model correctly estimates the category where a pair of species falls. However, the empirical fit provides reasonable p-values for the low-correlation cases. Even controlling for habitat structure (only considering pairwise correlation among species prevelant in the same habitat) this pattern persists. This indicates that the DPM-DMM is useful for qualitatively capturing species correlations but does not provide precise quantitative estimates across disparate niches.

English channel bacterial time series

Supplementary Figure 1 shows the ecological painting for the moderately filtered L4 time series data. This indicates moderate successional patterns over the year, beginning with two closely related ecostates in April, transitioning to two other states for the summer and finally transitioning back to the original states in the late autumn, consistent with the broader seasonality observed in this time series. This pattern is not a strongly defined seasonal transitions and possess multiple closely related states, as suggested in the original paper presenting the entire six-year time series from which the data were excerpted. Without additional context – more intensive temporal sampling, another spatial location, or envionmental covariates – it is uncertain if the painting reveals inconsistency in the model performance or genuine heterogeneity within the data, although reports on the larger data set indicate the latter.

Figure 3 summarizes the PPD analysis for L4 time series. The p-value for each species within each sample shows no systematic issues, except for poor performance across six low count samples (horizontal bands). The model appears to capture mean frequency for nearly all species, while systematically overestimating the amount of variation in counts across samples. The number of zeros within a sample are generally well-modeled, though a significant fraction of samples have overestimates of these numbers. As to the pairwise correlations across species, the model performs generally well for most pairs, although the model overestimates the correlation for a distinct fraction of most frequent species. For moderately frequent and infrequent species, the model shows excellent agreement with the data.

Gulf of Maine copepod time series

Figure 4 gives the ecological painting for the GoM copepod time series, with each panel organized from left to right by month and bottom to top by distance from New Bedford, Massachusetts. The painting shows strong seasonal and spatial trends, as frequently noted in the literature [48, 51, 60]. Both the mid-summer months (Jun.-Aug.) and mid-winter months (Nov.-Feb.) exhibit substantially less spatial variation than other months. The ecostates associated with these relatively quiescent periods show a single or small number of dominant species, with relatively high variation. The transitional ecostates between these periods exhibit more complex taxon composition.

The painting also highlights the dramatic shift in species that occurred from 1990-2001 relative to the years before and after [51, 27]. In the years outside the shift, the winter-summer cycle is fairly stable, with only a short transition period associated with other ecostates. This transition is largely dominated by changes in the frequency of *Calanus finmarchicus*, a crucial species in the north Atlantic food web. During this shift, this stability diminishes, with long transitions associated with higher entropy ecostates. During this period, neighboring samples in time and space are less likely to share an ecostate and novel, highly complex ecostates are more common, suggest ecological distress. These patterns are consistent with a major ecological disturbance, such as global warming or nutrient forcing.

Figure 5 shows inferred PPD for the dataset. Considering the species/samples values, the performance for highly abundant species is generally good, with poor performance for all other taxa (note the correspondence between p-value and mean; upper right and center panels). However, we see strong performance for the mean, standard deviation, zeros, and pairwise correlation. This is consistent with an ecosystem largely determined by a small number of highly abundant species that are likely not exchangeable (in an ecological sense) with other taxa, contrary to the the UNTB. Consequently, the model fits those species well but poorly captures the remaining species.

Discussion

This work investigates the performance of an infinite dimensional version of the DMM for the ecological count data. It shows how this approach – like other implementations of the DMM – can be used to organize and interpret underlying sample dyanmics with a straight-forward visualization while guarding against overinterpretation by examining the posterior predictive fit. While the model effectively distills many of the important patterns within the data inferred via other non-model-based approaches, such as PCA or MDS, it also shows that some caution should be exercised in these analyses and suggests certain practices may improve the reliability of field investigations.

Posterior checks, especially PPDs, can be an important bulwark in ensuring that these models are treated with appropriate skepticism [6]. For the DMMs, this requirement is not particularly burdensome as PPDs are easy to simulate and can be interpreted using familar p-values as a metric. Appropriately used, these measures give researchers and their readers increased assurance in the reported conclusions. We encounter several issues in the analysis of the datasets above that suggest useful guidance for field researchers. We organize these suggestions below under the following headings: identifying problems with samples; relevant scales of ecological fluctuation; niches as ecological outgroups; and departures from niche neutrality.

Identifying problems with samples

In our examples, PPDs consistently demonstrate identify problematic samples from the DMM's perspective that may suggest of experimental issues. In the cases we present, sample-specific deviations from model expectations may be due to low sample counts, mislabeled samples, unusual taxa present, or departures from the DMM's assumptions. PPDs for the DMM provide a convenient tool for identifying these samples, showing low p-value'bands' that are indicative of poor performance for a sample across a subset of species. In the L4 example, bands across all species indicated insufficient counts for reliable inference within certain samples. In the MPHM painting, the tongue ecostate is found in the feces habitat, a likely mislabeling. The ecological painting provides an easy means to identify these questionable samples.

Relevant scales of ecological fluctuation

By ecological fluctuation, we mean the changes in environmental conditions that precipitate shifts in species composition and consequently ecostate. The easiest scale of ecological fluctution to analyze is the case-control study. As highlighted in their analysis of twin obsesity [25], the association between DMM components and case/control status yields a straight-forward statistical approach. A similar approach was recently used to analyze the microbiomes of human cancers [43]. However, ecologists are often interested in more sampling across more heterogeneous patterns of spatial, temporal or environmental change. While there is some guidance about the level of read-depth required for DM inference [30], there is little guidance available for dealing with ecostate variation.

The analysis of time series datasets here makes clear that ecologists should prepare a sampling design sufficient resolution to capture ecostate transitions. In practice, this means sampling at a temporal resolution substantially higher than the expected scale of the transition itself. For instance, the MPHM data is compelling in large part because it shows the relative constancy of species abundance in time and space. However, this consistency is revealed precisely because the sampling occurs at a time interval substantially more frequent than the scale of ecostate variation. Similarly, the absence of intensive sampling renders the conclusion of seasonality in the L4 time series uncertain. A seasonal transition that would be anticipated for a temperate marine microbiome is on the order of 90 days (a single season). A minimum sampling regime should then be an order of magnitude less than that, or less than 9 days, about 60% of the two week average interval for the L4 study. In other cases, such as spatial or environmental gradients, a similar rule of thumb can be employed.

Niches as ecological outgroups

The specific examples we consider here strongly suggest that researchers consider using ‘outgroup’ sampling as important tool for contextualizing their analyses. The term outgroup in phylogenetic analysis refers to a species included in the collection for the purpose of parsing the relatedness among the study population [45]. In the context of ecological data, we mean a sample that has some related properties to the main study target but is sufficiently dissimilar to provide a backdrop for understanding the degree of variation in the study population. An outgroup does not need to fall entirely outside the study domain. For instance, in the MPHM, the separate niches (feces, saliva, skin) provide simultaneous outgroups for each other, allowing more confident interpretation. In the copepod data set, the extensive temporal and spatial sampling provides is akin to an outgroup, though a set of open ocean samples would be closer to ideal. The analysis of the L4 time series suffers from the lack of an outgroup, leaving the apparent seasonality in doubt (though reports from the complete, six year dataset indicate the robustness of this observation).

Departures from the niche neutrality

To some researchers, departures from the DMM model assumptions mean that the UNTB and analysis based on the DMM can be abandoned immediately. For researchers content to still use this analysis, these departures are still useful: as the data depart from model expectations, the structure of these deviations can reveal important aspects of the ecology. In the GoM copepod data set, the domination by *C. finmarchicus* represents a departure from underlying model assumptions as it cannot be ‘swapped out’ for any other copepod species (ie it is not ecologically equivalent. This is clearly shown in the PPDs for each species: the model performs well for highly abundant *C. finmarchicus* and closely engaged taxa, but poorly for all other taxa. In the well-studied copepod context, this does not provide new information. However, in a novel contexts this departure would be a useful insight into underlying ecological dynamics.

We note that these departures are qualitative (‘bad’ fit with the PPDs) rather than quantitative.

An important avenue for future research with the DMM is to develop precise statistical tests for departures from model expectations with ecological understandings of their consequence, as shown in [22].

Conclusions

The DMM and its variants give powerful tools for understanding ecological count data. However, these models possess a number of weaknesses that could be addressed in the next generation of models. Most obviously, these approaches do not consider the total number of model counts within a sample, a critical indicator of ecosystem function. Researchers might naturally account for this by excluding exceptionally high- or low-count samples, though this would be better addressed by including this variation at the level of the model, as is possible with a negative binomial process [72]. The DMM also does not account for correlation across samples, as could be done using Gaussian process priors or hidden Markov models to 'borrow strength' across samples.

Finally, we do not believe that using the DMM requires one to take a strong position on the UNTB: the underlying ecostates can be treated as phenomenological clusters, rather than theoretically precise metacommunity structures, and still provide analytic utility. As the use of the coalescent model in phylogenetics is not substantially diminished by the frequent observation of violations of its assumptions, the approach of the DMM and similar models promise the possibility of a connection to the broader UNTB framework while often giving a practical means of interpreting datasets independent of its operation.

Figures

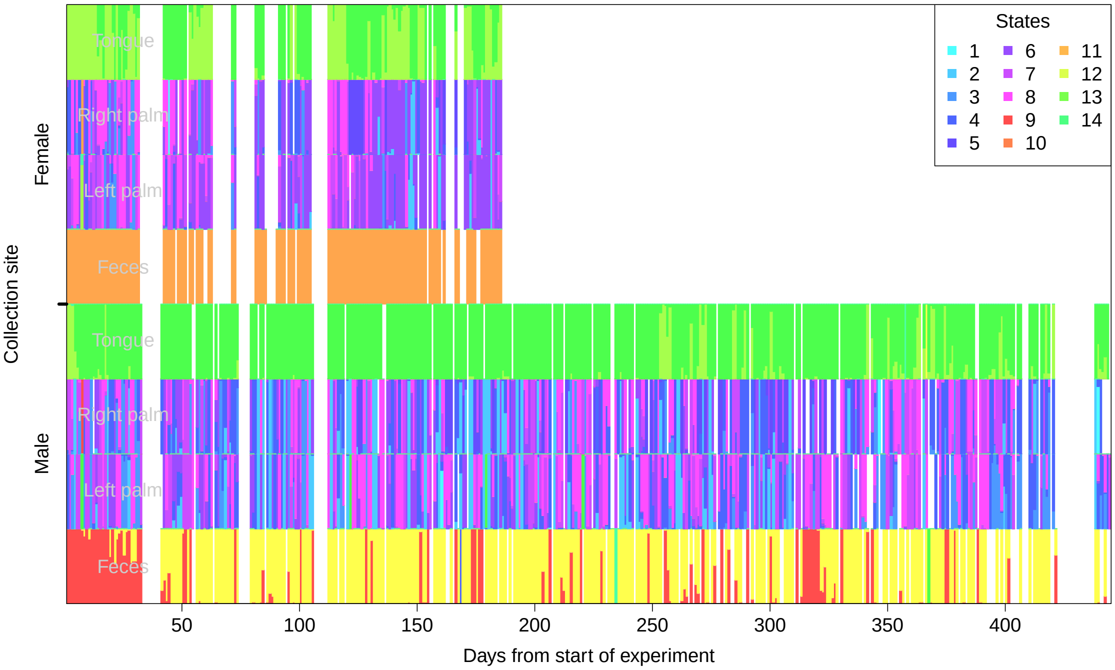

Figure 1: The ecological painting for the MPHM showing the ecostate for samples cloted across two individuals (female above, male below) and the four collection sites over approximately 450 days.

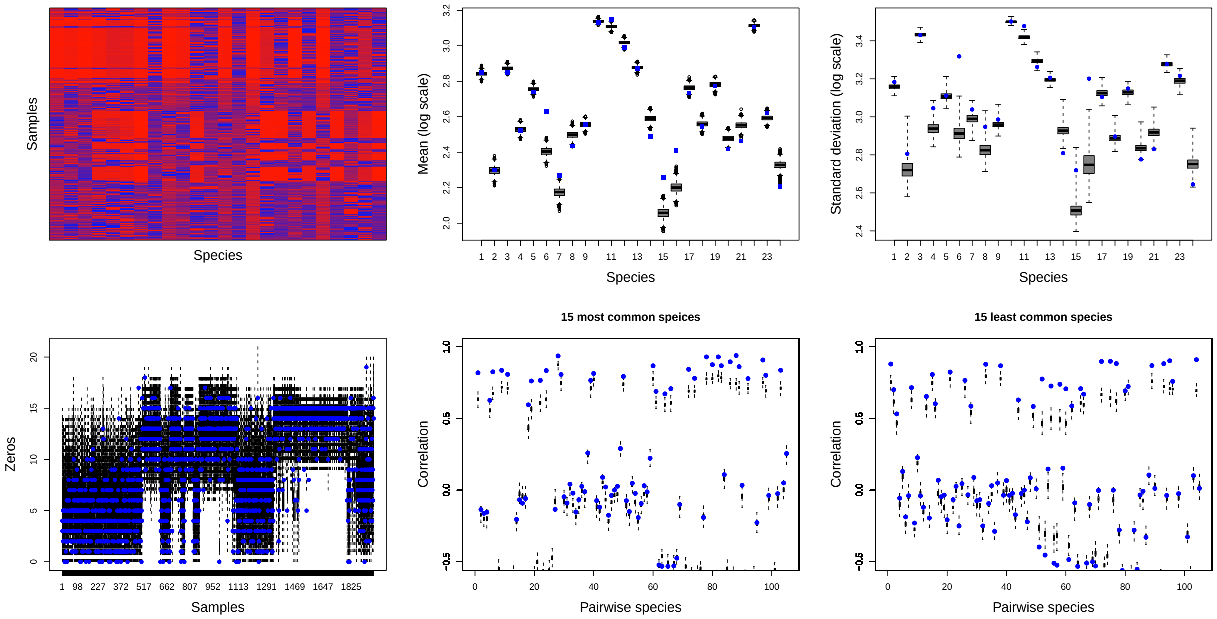

Figure 2: The PPP for the MPHM showing performance for each taxa within each sample (upper left), mean across all samples (upper middle), standard deviation across all samples (upper right), zeros within samples (lower left), pairwise correlation among abundant species (bottom middle) and infrequent species (bottom right). Boxplot show simulated values; blue dots show observed values

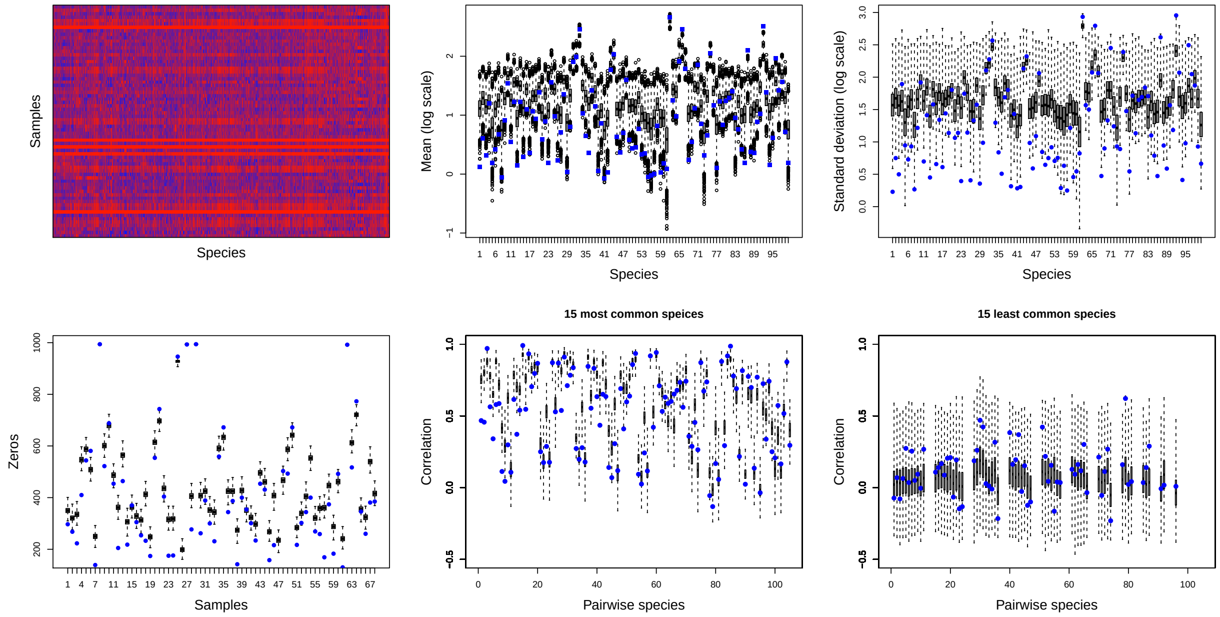

Figure 3: PPD summary for L4 amplicon time series.

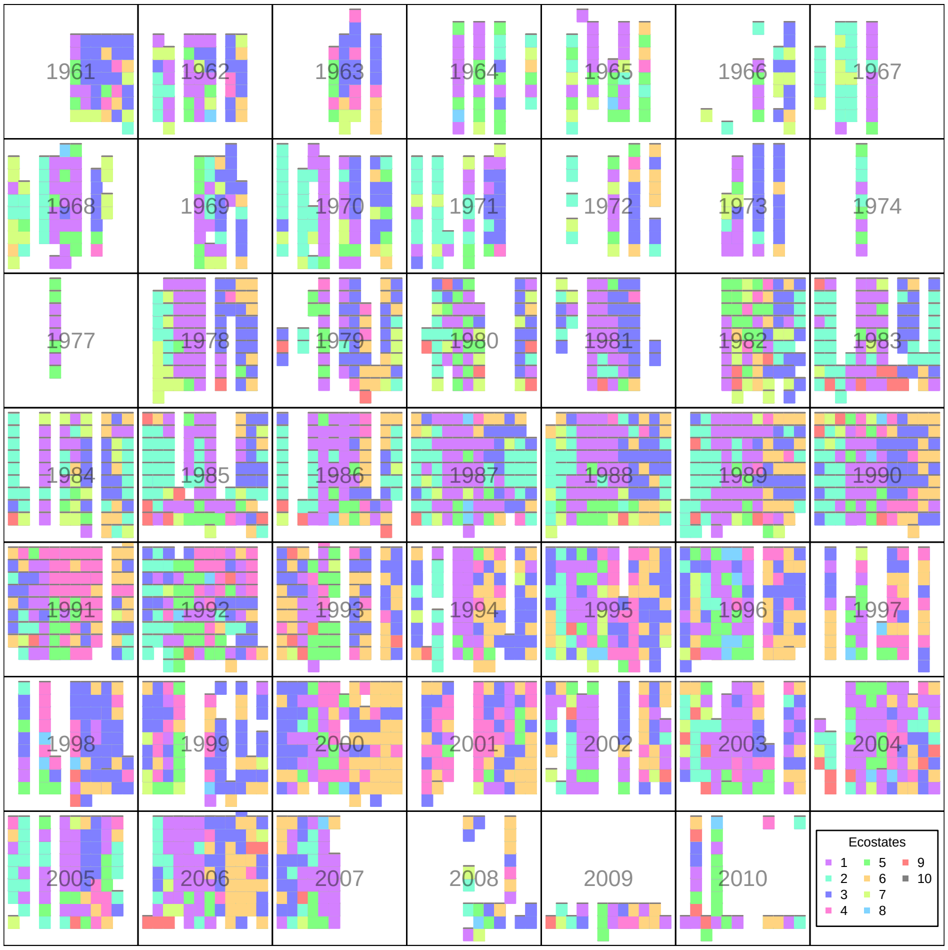

Figure 4: Ecological painting of GoM copepod data set, organized by year. Each panel corresponds to a year, with samples arranged spatially with Boston at the bottom and Yarmouth at the top.

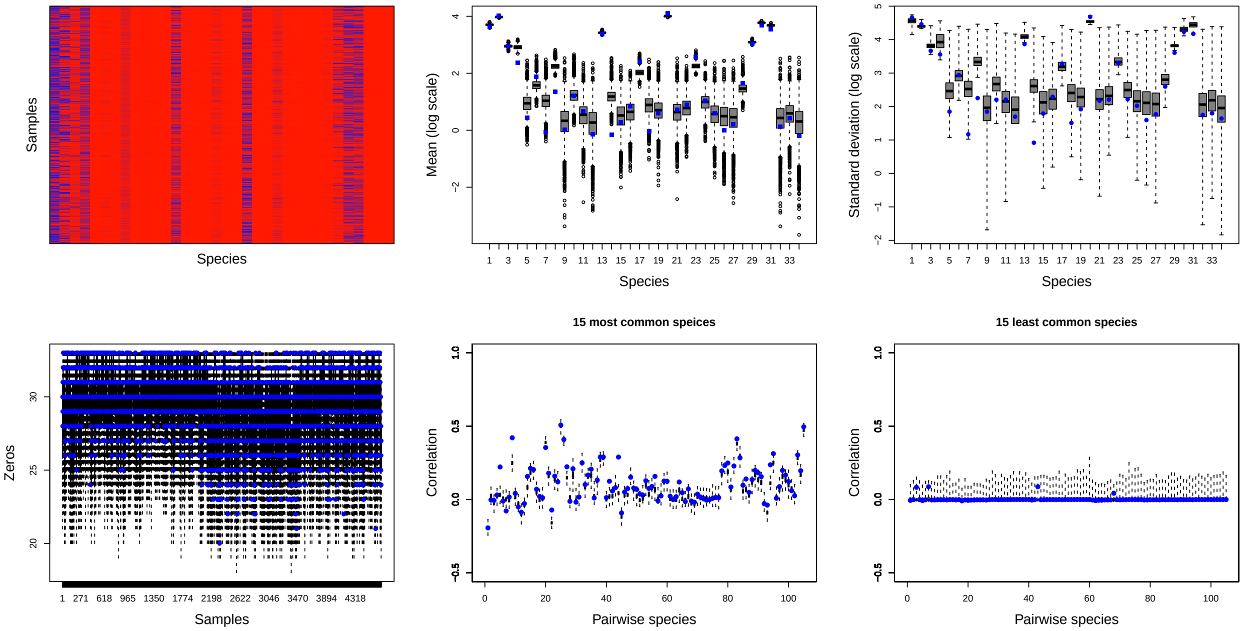

Figure 5: PPD summary for GOM copepod data set. Strong agreement for all aggregate metrics is opposed to performance by species within samples. Note the correspondence between high mean and good performance by species.

References

[1] Kjersti Aagaard, Kevin Riehle, Jun Ma, Nicola Segata, Toni-Ann Mistretta, Cristian Coarfa, Sabeen Raza, Sean Rosenbaum, Ignatia Van den Veyver, Aleksandar Milosavljevic, et al. A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy. *PloS ONE*, 7(6):e36466, 2012.

[2] Helen M Alexander, Bryan L Foster, Ford Ballantyne, Cathy D Collins, Janis Antonovics, and Robert D Holt. Metapopulations and metacommunities: combining spatial and temporal perspectives in plant ecology. *Journal of Ecology*, 100(1):88–103, 2012.

[3] Stefano Allesina and Mercedes Pascual. Network structure, predator–prey modules, and stability in large food webs. *Theoretical Ecology*, 1(1):55–64, 2008.

[4] David Alonso and Alan J McKane. Sampling hubbell's neutral theory of biodiversity. *Ecology Letters*, 7(10):901–910, 2004.

[5] Manimozhiyan Arumugam, Jeroen Raes, Eric Pelletier, Denis Le Paslier, Takuji Yamada, Daniel R Mende, Gabriel R Fernandes, Julien Tap, Thomas Bruls, Jean-Michel Batto, et al. Enterotypes of the human gut microbiome. *Nature*, 473(7346):174–180, 2011.

[6] Mark A Beaumont. Approximate bayesian computation in evolution and ecology. *Annual review of ecology, evolution, and systematics*, 41:379–406, 2010.

[7] Surojit Biswas, Meredith McDonald, Derek S Lundberg, Jeffery L Dangl, and Vladimir Jojic. Learning microbial interaction networks from metagenomic count data. In *Research in Computational Molecular Biology*, pages 32–43. Springer, 2015.

[8] James H Brown, James F Gillooly, Andrew P Allen, Van M Savage, and Geoffrey B West. Toward a metabolic theory of ecology. *Ecology*, 85(7):1771–1789, 2004.

[9] MG Bulmer. On fitting the poisson lognormal distribution to species-abundance data. *Biometrics*, pages 101–110, 1974.

[10] J Gregory Caporaso, Christian L Lauber, Elizabeth K Costello, Donna Berg-Lyons, Antonio Gonzalez, Jesse Stombaugh, Dan Knights, Pawel Gajer, Jacques Ravel, Noah Fierer, et al. Moving pictures of the human microbiome. *Genome Biol*, 12(5):R50, 2011.

[11] J Gregory Caporaso, Konrad Paszkiewicz, Dawn Field, Rob Knight, and Jack A Gilbert. The Western English Channel contains a persistent microbial seed bank. *The ISME journal*, 6(6):1089–1093, 2012.

[12] Emily S Charlson, Kyle Bittinger, Jun Chen, Joshua M Diamond, Hongzhe Li, Ronald G Collman, and Frederic D Bushman. Assessing bacterial populations in the lung by replicate analysis of samples from the upper and lower respiratory tracts. *PloS one*, 7(9):e42786, 2012.

[13] Jun Chen, Kyle Bittinger, Emily S Charlson, Christian Hoffmann, James Lewis, Gary D Wu, Ronald G Collman, Frederic D Bushman, and Hongzhe Li. Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics*, 28(16):2106–2113, 2012.

[14] Jun Chen and Hongzhe Li. Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *The Annals of Applied Statistics*, 7(1):418–442, 2013.

[15] Tom O Delmont, Emmanuel Prestat, Kevin P Keegan, Michael Faubladier, Patrick Robe, Ian M Clark, Eric Pelletier, Penny R Hirsch, Folker Meyer, Jack A Gilbert, et al. Structure, fluctuation and magnitude of a natural grassland soil metagenome. *The ISME journal*, 6(9):1677–1687, 2012.

[16] Maria Dornelas, Sean R Connolly, and Terence P Hughes. Coral reef diversity refutes the neutral theory of biodiversity. *Nature*, 440(7080):80–82, 2006.

[17] Alexei J Drummond and Andrew Rambaut. Beast: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology*, 7(1):214, 2007.

[18] Rampal S Etienne and Han Olff. Confronting different models of community structure to species-abundance data: a bayesian model comparison. *Ecology letters*, 8(5):493–504, 2005.

[19] Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, 6(4):733–760, 1996.

[20] Jack Gilbert, Folker Meyer, Lynn Schriml, Ian Joint, Martin Muhling, and Dawn Field. Metagenomes and metatranscriptomes from the l4 long-term coastal monitoring station in the Western English Channel. *Standards in genomic sciences*, 3(2):183–193, 2010.

[21] Jack A Gilbert, Simon Thomas, Natalie A Cooley, Anna Kulakova, Dawn Field, Tim Booth, John W McGrath, John P Quinn, and Ian Joint. Potential for phosphonoacetate utilization by marine bacteria in temperate coastal waters. *Environmental microbiology*, 11(1):111–125, 2009.

[22] Keith Harris, Todd L Parsons, Umer Z Ijaz, Leo Lahti, Ian Holmes, and Christopher Quince. Linking statistical and ecological theory: Hubbell's unified neutral theory of biodiversity as a hierarchical dirichlet process. *arXiv preprint arXiv:1410.4038*, 2014.

[23] Garrett Hellenthal, Adam Auton, and Daniel Falush. Inferring human colonization history using a copying model. *PLoS Genetics*, -:10–1371, 2008.

[24] J Hirai, M Kuriyama, T Ichikawa, K Hidaka, and A Tsuda. A metagenetic approach for revealing community structure of marine planktonic copepods. *Molecular ecology resources*, 15(1):68–80, 2015.

[25] Ian Holmes, Keith Harris, and Christopher Quince. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS One*, 7(2):e30126, 2012.

[26] Stephen P Hubbell. *The unified neutral theory of biodiversity and biogeography (MPB-32)*, volume 32. Princeton University Press, 2001.

[27] Catherine L Johnson, Jeffrey A Runge, K Alexandra Curtis, Edward G Durbin, Jonathan A Hare, Lewis S Incze, Jason S Link, Gary D Melvin, Todd D OBrien, and Lou Van Guelpen. Biodiversity and ecosystem function in the gulf of maine: pattern and role of zooplankton and pelagic nekton. *PLoS One*, 6(1):e16491, 2011.

[28] Axel Kleidon, Yadvinder Malhi, and Peter M Cox. Maximum entropy production in environmental and ecological systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1545):1297–1302, 2010.

[29] Jeremy E Koenig, Aymé Spor, Nicholas Scalfone, Ashwana D Fricker, Jesse Stombaugh, Rob Knight, Largus T Angenent, and Ruth E Ley. Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4578–4585, 2011.

[30] Patricio S La Rosa, J Paul Brooks, Elena Deych, Edward L Boone, David J Edwards, Qin Wang, Erica Sodergren, George Weinstock, and William D Shannon. Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PloS one*, 7(12):e52078, 2012.

[31] Steven J Lade, Alessandro Tavoni, Simon A Levin, and Maja Schlüter. Regime shifts in a social-ecological system. *Theoretical ecology*, 6(3):359–372, 2013.

[32] Mathew A Leibold, M Holyoak, N Mouquet, P Amarasekare, JM Chase, MF Hoopes, RD Holt, JB Shurin, R Law, D Tilman, et al. The metacommunity concept: a framework for multi-scale community ecology. *Ecology letters*, 7(7):601–613, 2004.

[33] Penelope K Lindeque, Helen E Parry, Rachel A Harmer, Paul J Somerfield, and Angus Atkinson. Next generation sequencing reveals the hidden diversity of zooplankton assemblages. *PLoS ONE*, DOI: 10.1371:journal.pone.0081327, 2013.

[34] Elena Litchman and Christopher A Klausmeier. Trait-based community ecology of phytoplankton. *Annual Review of Ecology, Evolution, and Systematics*, pages 615–639, 2008.

[35] Zhenqiu Liu, Fengzhu Sun, Jonathan Braun, Dermot PB McGovern, and Steven Piantadosi. Multilevel regularized regression for simultaneous taxa selection and network construction with metagenomic count data. *Bioinformatics*, 31(7):1067–1074, 2015.

[36] Catherine Lozupone and Rob Knight. Unifrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology*, 71(12):8228–8235, 2005.

[37] Robert McCredie May. *Stability and complexity in model ecosystems*, volume 6. Princeton University Press, 2001.

[38] Brian J McGill. A test of the unified neutral theory of biodiversity. *Nature*, 422(6934):881–885, 2003.

[39] Xiao-Li Meng. Posterior predictive p-values. *The Annals of Statistics*, pages 1142–1160, 1994.

[40] Folker Meyer, Daniel Paarmann, Mark D'Souza, Robert Olson, Elizabeth M Glass, Michael Kubal, Tobias Paczian, A Rodriguez, Rick Stevens, Andreas Wilke, et al. The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9(1):386, 2008.

[41] A Mougi and M Kondoh. Diversity of interaction types and ecological community stability. *Science*, 337(6092):349–351, 2012.

[42] Peter Müller and Fernando A Quintana. Nonparametric Bayesian data analysis. *Statistical science*, pages 95–110, 2004.

[43] Geicho Nakatsu, Xiangchun Li, Haokui Zhou, Jianqiu Sheng, Sunny Hei Wong, William Ka Kai Wu, Siew Chien Ng, Ho Tsoi, Yujuan Dong, Ning Zhang, et al. Gut mucosal microbiome across stages of colorectal carcinogenesis. *Nature communications*, 6, 2015.

[44] Radford M Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

[45] Kevin C Nixon and James M Carpenter. On outgroups. *Cladistics*, 9(4):413–426, 1993.

[46] TD OBrien. Copepod: A global plankton database. *NOAA Technical Memorandum NMFS-F/SPO-73*, page 19, 2005.

[47] Michael L Pace, Jonathan J Cole, Stephen R Carpenter, and James F Kitchell. Trophic cascades revealed in diverse ecosystems. *Trends in ecology & evolution*, 14(12):483–488, 1999.

[48] Andrew J Pershing, Charles H Greene, Jack W Jossi, Loretta O'Brien, Jon KT Brodziak, and Barbara A Bailey. Interdecadal variability in the Gulf of Maine zooplankton community, with potential impacts on fish recruitment. *ICES Journal of Marine Science: Journal du Conseil*, 62(7):1511–1523, 2005.

[49] Owen L Petchey, Andrew P Beckerman, Jens O Riede, and Philip H Warren. Size, foraging, and food web structure. *Proceedings of the National Academy of Sciences*, 105(11):4191–4196, 2008.

[50] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic acids research*, page gks1219, 2012.

[51] Nicholas R Record, Andrew J Pershing, and Jack W Jossi. Biodiversity as a dynamic variable in the gulf of maine continuous plankton recorder transect. *Journal of plankton research*, 32(12):1675–1684, 2010.

[52] Robert E Ricklefs. The unified neutral theory of biodiversity: do the numbers add up? *Ecology*, 87(6):1424–1431, 2006.

[53] Christian S Riesenfeld, Patrick D Schloss, and Jo Handelsman. Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.*, 38:525–552, 2004.

[54] Christian Ritter and Martin A Tanner. Facilitating the Gibbs sampler: the Gibbs stopper and the griddy-Gibbs sampler. *Journal of the American Statistical Association*, 87(419):861–868, 1992.

[55] Fredrik Ronquist and John P Huelsenbeck. Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574, 2003.

[56] Noah A Rosenberg and Magnus Nordborg. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, 3(5):380–390, 2002.

[57] James Rosindell, Stephen P Hubbell, and Rampal S Etienne. The unified neutral theory of biodiversity and biogeography at age ten. *Trends in ecology & evolution*, 26(7):340–348, 2011.

[58] Patrick D Schloss, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan A Lesniewski, Brian B Oakley, Donovan H Parks, Courtney J Robinson, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, 75(23):7537–7541, 2009.

[59] Mahdi Shafiei, Katherine A Dunn, Eva Boon, Shelley M MacDonald, David A Walsh, Hong Gu, and Joseph P Bielawski. Biomico: a supervised bayesian model for inference of microbial community structure. *Microbiome*, 3(1):1–15, 2015.

[60] Karen Stamieszkin, Andrew J Pershing, Nicholas R Record, Cynthia H Pilskaln, Hans G Dam, and Leah R Feinberg. Size as the master trait in modeled copepod fecal pellet carbon flux. *Limnology and Oceanography*, 60(6):2090–2107, 2015.

[61] Helen L Steele and Wolfgang R Streit. Metagenomics: advances in ecology and biotechnology. *FEMS Microbiology Letters*, 247(2):105–111, 2005.

[62] Nathan M Stein and Xiao-Li Meng. Practical perfect sampling using composite bounding chains: the Dirichlet-multinomial model. *Biometrika*, 100(4):817–830, 2013.

[63] Matthew Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

[64] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476), 2006.

[65] Susannah Green Tringe, Christian Von Mering, Arthur Kobayashi, Asaf A Salamov, Kevin Chen, Hwai W Chang, Mircea Podar, Jay M Short, Eric J Mathur, John C Detter, et al. Comparative metagenomics of microbial communities. *Science*, 308(5721):554–557, 2005.

[66] Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire M Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. The human microbiome project. *Nature*, 449(7164):804–810, 2007.

[67] John Wakeley. *Coalescent theory: an introduction*, volume 1. Roberts & Company Publishers Greenwood Village, Colorado, 2009.

[68] Dan L Warren and Stephanie N Seifert. Ecological niche modeling in maxent: the importance of model complexity and the performance of model selection criteria. *Ecological Applications*, 21(2):335–342, 2011.

[69] Gary D Wu, Jun Chen, Christian Hoffmann, Kyle Bittinger, Ying-Yu Chen, Sue A Keilbaugh, Meenakshi Bewtra, Dan Knights, William A Walters, Rob Knight, et al. Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105–108, 2011.

[70] Fan Xia, Jun Chen, Wing Kam Fung, and Hongzhe Li. A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*, 69(4):1053–1063, 2013.

[71] Xiao Xiao, Daniel J McGlinn, and Ethan P White. A strong test of the maximum entropy theory of ecology. *The American Naturalist*, 185(3):E70–E80, 2015.

[72] MengChu Zhou and Lawrence Carin. Negative binomial process count and mixture modeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(2):307–320, 2015.

[73] Yanjiao Zhou, Hongyu Gao, Kathie A Mihindukulasuriya, Patricio S La Rosa, Kristine M Wylie, Tatiana Vishnivetskaya, Mircea Podar, Barb Warner, Phillip I Tarr, David E Nelson, et al. Biogeography of the ecosystems of the healthy human body. *Genome Biol*, 14(1):R1, 2013.