JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 29, Number 0, 2022

Pp. 1–14
DOI: 10.1089/cmb.2021.0595

Waterman Special Issue

Variational Approximation-Based Model Selection for Microbial Network Inference

SHIBU YOOSEPH and SAHAR TAVAKOLI

ABSTRACT

Microbial associations are characterized by both direct and indirect interactions between the constituent taxa in a microbial community, and play an important role in determining the structure, organization, and function of the community. Microbial associations can be represented using a weighted graph (microbial network), whose nodes represent taxa and edges represent pairwise associations. A microbial network is typically inferred from a sample–taxa matrix that is obtained by sequencing multiple biological samples and identifying the taxa counts in each sample. However, it is known that microbial associations are impacted by environmental and/or host factors. Thus, a sample–taxa matrix generated in a microbiome study involving a wide range of values for the environmental and/or clinical metadata variables may in fact be associated with more than one microbial network. In this study, we consider the problem of inferring multiple microbial networks from a given sample–taxa count matrix.

Each sample is a count vector assumed to be generated by a mixture model consisting of component distributions that are multivariate Poisson log-normal. We present a variational expectation maximization algorithm for the model selection problem to infer the correct number of components of this mixture model. Our approach involves reframing the mixture model as a latent variable model, treating only the mixing coefficients as parameters, and subsequently approximating the marginal likelihood using an evidence lower bound framework. Our algorithm is evaluated on a large simulated dataset generated using a collection of different graph structures (band, hub, cluster, random, and scale-free).

Keywords: microbiome, mixture models, networks, variational approximation.

1. INTRODUCTION

THE STRUCTURE OF A MICROBIAL COMMUNITY and the organization of its constituent members (taxa) are determined by a combination of their mutual interactions and other factors such as the availability of carbon sources, energy, and nutrients, and the characteristics of the surrounding environment (Falkowski

Department of Computer Science, Genomics and Bioinformatics Cluster, University of Central Florida, Orlando, Florida, USA.

et al., 2008; Hibbing et al., 2010; Williamson and Yooseph, 2012). These variables determine whether the community contains only a handful of taxa with very little strain variation [such as in limiting environments like acid mine drainage (Tyson et al., 2004)] or whether they contain a moderate-to-high number of taxa with a large number of variants [such as in the human environment (Methe et al., 2012), oligotrophic open oceans (Rusch et al., 2007), and nutrient-rich soils (Vogel et al., 2009)].

Microbial community composition can be obtained by sequencing the DNA extracted from a biological sample collected from the environment of interest. Microbiome sequence data are generated either using a targeted approach, involving the sequencing of a taxonomic marker gene [for instance, the 16S ribosomal RNA gene, which is found in all bacteria (Woese and Fox, 1977)] or using a whole-genome shotgun sequencing approach (Venter et al., 2004); the latter approach can be used to deduce both the taxonomic composition and the functional potential of the community.

Here, we study the computational problem of inferring microbial associations from microbiome data. We use the term *microbial association* to capture both influences and interactions between microbial taxa. In a microbial community, the presence and abundance of one taxonomic group may either *directly* or *indirectly* influence the abundance of another taxonomic group (Hibbing et al., 2010). For instance, two microbial taxa may directly influence each other through interactions involving exchange of metabolites or other products, or by competing for the same resources. Alternately, two microbial taxa may not directly communicate or compete for the same resources, but instead one taxon could interact with other members of the community, and these interactions could indirectly influence resource availability for the other taxon. Information about associations between taxa can provide important insights into the ecology of the microbial community.

Microbial associations can be represented using a weighted graph (*microbial network*) whose nodes represent taxa and undirected edges between nodes represent associations. Edge weights capture the strength of the associations, and the edge weight sign reflects whether the association is positive or negative (Layeghifard et al., 2017). This graph representation can be used to model a variety of microbial interactions, including competition and cooperation (Loftus et al., 2021). Microbial associations can be inferred from the underlying covariance structure of the community, which can be calculated using taxa abundances. The covariance matrix is estimated from a sample–taxa count matrix; this count matrix is generated by sequencing biological samples collected from the environment of interest and identifying the counts of taxa in each sample.

Typically, the study of a microbial community in a particular environment assumes a *single* covariance structure, and computational methods have been developed to address this estimation problem (Layeghifard et al., 2017), including approaches based on probabilistic graphical models (Kurtz et al., 2015; Biswas et al., 2016; Loftus et al., 2021) and on latent variable models (Friedman and Alm, 2012; Fang et al., 2015). However, with the use of high-throughput next-generation DNA sequencing technologies (Quail et al., 2012) that allow for cost-effectively obtaining data from biological samples, microbiome studies now routinely collect, and generate data from a large number of samples. In these situations, a microbiome study involving a large cohort or including a wide range of metadata variables (environmental and/or clinical) may in fact be sampling from a community where the microbial associations between taxa are not the same across all intervals of the metadata values. In other words, the microbiome samples in the study may be associated with more than one underlying covariance structure (and thus, more than one microbial network).

Motivated by this scenario, we recently developed an extension to the single network inference problem. In this extension, we treat the inference problem in a mixture model framework based on generative models (Tavakoli and Yooseph, 2019) and solve the following computational problem: given a sample–taxa count matrix generated by a mixture model with K component distributions, estimate the mixing coefficients and the parameters of the K component distributions. The component distributions model taxa count data, and each component is associated with one precision matrix (and thus, one microbial network). In our framework, referred to as MixMPLN (Tavakoli and Yooseph, 2019), we assume that the taxa counts are generated by multivariate Poisson log-normal (MPLN) distributions (Aitchison and Ho, 1989; Inouye et al., 2017). We estimate the parameters of the MixMPLN model in a maximum likelihood setting using an optimization technique based on the minorization–maximization principle (Lange, 2016).

We note that the MPLN distribution has been used previously for the single network inference problem (Biswas et al., 2016; Chiquet et al., 2019). While distributions like the multinomial or the Dirichlet–Multinomial have been popular choices for modeling microbial count data in certain situations (Holmes

et al., 2012; La Rosa et al., 2012), these distributions cannot capture both positive and negative associations between taxa. On the other hand, the MPLN distribution can be used to model multivariate count data and its covariance matrix can capture both types of microbial associations.

While a mixture model framework can be used to study the multiple network scenario, in practice however, we do not have a priori knowledge of the value of K, the number of component distributions. In this article, we propose a variational approximation algorithm to determine the correct value of K for the MixMPLN framework, and solve the model selection problem in a statistically principled manner. As part of our approach, we reformulate the mixture model as a latent variable model (Corduneanu and Bishop, 2001), and treat only the mixing coefficients as parameters, while treating all other variables, including the means and precision matrices of the component distributions, as latent variables. We use suitable factor distributions involving the latent variables and provide a variational expectation maximization (EM) algorithm to compute the parameters of these factor distributions to approximate the true marginal likelihood. We evaluate our approach using simulated sample–taxa count matrices generated using different classes of microbial network graph structures.

2. METHODS

Notation: Given a matrix X, we use $X_{:i}$ to denote its ith column, $X_{j:}$ to denote its jth row, and x_{ji} to denote its entry in row j and column i. We use n to denote the number of samples, d to denote the number of taxa, and K to denote the number of mixture components. Unless otherwise specified, all vectors are assumed to be column vectors. In the equations below, we associate the variables i, j, and l with samples, taxa, and mixture components, respectively.

2.1. The MPLN distribution

The MPLN distribution can be used to model count data (Aitchison and Ho, 1989). It has parameters μ and Ω, where μ is the d-dimensional mean vector and $\Omega_{d\times d}$ is the precision matrix of the distribution. A sample $A=(a_1, \ldots, a_d)^T$ generated by this distribution is a d-dimensional count vector with the following property:

$$\begin{aligned} a_j|\lambda_j &\sim \mathbb{P}(e^{\lambda_j}) \\ (\lambda_1, \ldots, \lambda_d)^T &\sim \mathbb{N}_d(\mu, \Omega) \end{aligned} \tag{1}$$

where $\mathbb{P}(c)$ denotes a Poisson distribution with mean c, and $\mathbb{N}_d(\mu, \Omega)$ denotes a d-dimensional multivariate Gaussian distribution with mean μ and precision matrix Ω. That is, an MPLN distribution has two layers, with the observed count vector (i.e., sample) being generated by a mixture of independent Poisson distributions whose means are latent (or hidden), such that the logarithm of the Poisson means follows a multivariate Gaussian distribution. We use $\lambda = (\lambda_1, \lambda_2, .. \lambda_d)^T$ to denote the latent variable vector representing the logarithm of the Poisson means that is associated with the sample A.

The probability density function $p(A|\mu, \Omega)$ of the MPLN distribution with parameters μ and Ω can be written as (Aitchison and Ho, 1989):

$$p(A|\mu, \Omega) = \int_{\mathbb{R}^d} p(A, \lambda|\mu, \Omega)\, \mathrm{d}\lambda \tag{2}$$

where,

$$p(A, \lambda|\mu, \Omega) = \left[\prod_{j=1}^{d} \frac{e^{-e^{\lambda_j}} e^{\lambda_j a_j}}{a_j!}\right] (2\pi)^{-d/2} |\Omega|^{1/2} e^{-\frac{1}{2}[\lambda-\mu]^T \Omega [\lambda-\mu]} \tag{3}$$

and $|\Omega|$ denotes the determinant of Ω. No simplification of the integral in Equation (2) is known.

2.2. Mixture of MPLN distributions

The probability of a sample $A = (a_1, \ldots, a_d)^T$ generated by a mixture model with K MPLN component distributions and a mixing coefficient vector $\phi = (\phi_1, \phi_2, .., \phi_K)$ can be written as

$$p(A|\phi,\ \mu_{(1)},\ \Omega_{(1)},\ ..,\ \mu_{(K)},\ \Omega_{(K)}) = \sum_{l=1}^{K} \phi_l p(A|\mu_{(l)},\ \Omega_{(l)}) = \sum_{l=1}^{K} \phi_l [\int p(A,\ \lambda_{(l)}|\mu_{(l)},\ \Omega_{(l)})\ \mathrm{d}\lambda_{(l)}] \tag{4}$$

where $\sum_{l=1}^{K}\ \phi_l = 1$. In Equation (4), the variables $\mu_{(l)}$ and $\Omega_{(l)}$ denote the mean vector and precision matrix, respectively, of the lth component distribution, and $\lambda_{(l)}$ denotes the latent variable vector of sample A that is associated with the lth component.

Now, let $X_{d\times n}$ denote a sample–taxa count matrix with d taxa and n samples that is generated by mixture of K MPLN distributions. Also, let $\Lambda_{(l)}$ denote the $d\times n$ matrix of latent variable vectors of the n samples that is associated with the lth component. That is, column vector $\Lambda_{(l):i}$ is associated with sample $X_{:i}$. We also use λ_{lji} to denote the jth entry in column vector $\Lambda_{(l):i}$. Then, the probability of the observed sample–taxa count matrix X, given the parameters of the mixture model, can be written as

$$p(X|\phi,\ \mu_{(1)},\ \Omega_{(1)},\ ..,\ \mu_{(K)},\ \Omega_{(K)}) = \prod_{i=1}^{n}\sum_{l=1}^{K} \phi_l p(X_{:i}|\mu_{(l)},\ \Omega_{(l)}) = \prod_{i=1}^{n}\sum_{l=1}^{K} \phi_l \left[\int p(X_{:i},\ \Lambda_{(l):i}|\mu_{(l)},\ \Omega_{(l)})\ \mathrm{d}\Lambda_{(l):i}\right] \tag{5}$$

where

$$p(X_{:i},\ \Lambda_{(l):i}|\Theta_{(l)}) = \left[\prod_{j=1}^{d} \frac{e^{-e^{\lambda_{lji}}} e^{\lambda_{lji}x_{ji}}}{x_{ji}!}\right] (2\pi)^{-d/2} |\Omega_{(l)}|^{1/2} e^{-\frac{1}{2}\left[\Lambda_{(l):i}\ -\ \mu_{(l)}\right]^T \Omega_{(l)} \left[\Lambda_{(l):i}\ -\ \mu_{(l)}\right]} \tag{6}$$

This mixture model is associated with K microbial networks, where the lth network ($1\ \leq\ l\ \leq\ K$) has adjacency matrix equal to the precision matrix $\Omega_{(l)}$.

2.3. The latent variable model

We reformulate the mixture model given in Equation (5) as a latent variable model (Bishop, 2006), in which we treat only the mixing coefficients ϕ_l's as parameters while all other variables, including $\Lambda_{(l)}$, $\mu_{(l)}$, and $\Omega_{(l)}$, where $1\ \leq\ l\ \leq\ K$, are treated as latent variables.

Let $\Theta = \mathcal{L}\cup\mathcal{M}\cup\mathcal{T}\cup\mathcal{S}$ denote the set of all latent variables in our model, where $\mathcal{L}=\{\Lambda_{(l)}|1\leq l\leq K\}$, $\mathcal{M} = \{\mu_{(l)}|1\leq l\leq K\}$, $\mathcal{T} = \{\Omega_{(l)}|1\leq l\leq K\}$, and $\mathcal{S}=\{S_i|1\leq i\leq n\}$. The set $\mathcal{S}$ denotes component membership information for samples, where $S_i=(s_{i1},\ \ldots,\ s_{iK})^T$ is a K-dimensional binary vector, also called a 1-of-K binary vector (Bishop, 2006) that is associated with sample $X_{:i}$. This vector has the property that if $X_{:i}$ was generated by component r then $s_{ir}=1$, and that $s_{il}=0$, for all $l\neq r$.

We now describe the different parts of the generative model. Each S_i is drawn from a multinomial distribution; that is, $S_i\sim$Multinomial$(1,\ \phi)$. We have that

$$p(\mathcal{S}|\phi) = \prod_{i=1}^{n} p(S_i|\phi) = \prod_{i=1}^{n}\prod_{l=1}^{K} \phi_l^{s_{il}}$$

Conditional on $\mathcal{S}$, each sample is assumed to be independently drawn from an MPLN distribution with parameters $\mu_{(l)}$ and $\Omega_{(l)}$. Upon selection of component l, both sample $X_{:i}$ and its associated latent variable vector $\Lambda_{(l):i}$ are generated. Thus,

$$p(X,\mathcal{L}|\mathcal{M},\mathcal{T},\mathcal{S}) = \prod_{i=1}^{n}\prod_{l=1}^{K} p(X_{:i},\ \Lambda_{(l):i}|\mu_{(l)},\ \Omega_{(l)})^{s_{il}}$$

Marginalizing the function $p(X,\mathcal{L}|\mathcal{M},\mathcal{T},\mathcal{S})\times p(\mathcal{S}|\phi)$ over $\mathcal{S}$ and $\mathcal{L}$ results in Equation (5). We also introduce conjugate priors over each $\mu_{(l)}$ and $\Omega_{(l)}$. We assume that $\mu_{(l)}\sim\mathbb{N}_d(0,\beta I)$ where I is the $d\times d$ identity matrix, and β is a fixed parameter. We also assume that $\Omega_l\sim\mathbb{W}(\nu,\ V)$, where $\mathbb{W}(\nu,\ V)$ is the Wishart distribution (Wishart, 1928) with fixed degrees of distribution ν and fixed scale matrix V. The density function for the Wishart distribution is given as

$$\mathbb{W}(\Omega|v,\ V) = \frac{|V|^{v/2}|\Omega|^{(v-d-1)/2}}{2^{vd/2}\pi^{d(d-1)/4}\prod_{j=1}^{d}\Gamma\left(\frac{v+1-j}{2}\right)}\exp\left(-\frac{1}{2}\mathrm{Tr}(V\Omega)\right)$$

where Tr(.) and $\Gamma(.)$ denote the matrix trace function and Gamma function, respectively.

Finally, we set $p(\mathcal{M}) = \prod_{l=1}^{K} p(\mu_{(l)})$ and $p(\mathcal{T}) = \prod_{l=1}^{K} p(\Omega_{(l)})$. Taken together, these specifications allow us to describe the joint distribution of X and all latent variables, conditioned on the mixing coefficients, as

$$\begin{aligned} p(X,\ \Theta|\phi) &= p(X,\ \mathcal{L},\ \mathcal{M},\ \mathcal{T},\ \mathcal{S}|\phi) \\ &= p(X,\ \mathcal{L}|\mathcal{M},\ \mathcal{T},\ \mathcal{S}) \times p(\mathcal{S}|\phi) \times p(\mathcal{M}) \times p(\mathcal{T}) \end{aligned} \tag{7}$$

2.4. The evidence lower bound function

The marginal likelihood function $p(X|\phi)$ can be obtained by integrating Equation (7) over all latent variables in Θ. However, the function defined in this manner is not analytically tractable. Instead, we employ a variational approximation method involving a lower bound on the marginal log-likelihood function. This lower bound, called the Evidence Lower Bound (ELBO) function (Bishop, 2006; Tzikas et al., 2008), will then be maximized with respect to the mixing coefficients. The ELBO function $Q(\Theta)$ is defined as

$$\begin{aligned} ELBO(Q) &= \int Q(\Theta)\log\left[\frac{p(X,\ \Theta|\phi)}{Q(\Theta)}\right]\mathrm{d}\Theta \\ &= \langle \log p(X,\ \Theta|\phi)\rangle_{\Theta} - \langle \log Q(\Theta)\rangle_{\Theta} \end{aligned} \tag{8}$$

where $\langle . \rangle_{\Theta}$ denotes the expectation over the distribution $Q(\Theta)$. For any function $Q(\Theta)$, the following identity holds (Bishop, 2006):

$$ELBO(Q) \leq \log\left(\int p(X,\ \Theta|\phi)\mathrm{d}\Theta\right) = \log p(X|\phi)$$

Our goal is to maximize $ELBO(Q)$ using some choice of $Q(\Theta)$. The difference between $ELBO(Q)$ and $\log p(X|\phi)$ can be shown to be the Kullback–Leibler (KL) distance between $Q(\Theta)$ and the posterior distribution $p(\Theta|X,\ \phi)$. Thus, $ELBO(Q)$ is maximum when $Q(\Theta)$ is equal to the posterior (Bishop, 2006). Let $\Theta = \mathcal{L} \cup \mathcal{M} \cup \mathcal{T} \cup \mathcal{S} = \{\theta_i\}$. We assume that $Q(\Theta) = \prod_t q(\theta_t)$, that is, $Q(\Theta)$ is the product of independent factor distributions $q(\theta_t)$. With this assumption (Parisi, 1988), the form of the optimal factor distributions that minimize the KL distance can be computed (Bishop, 2006). For each t, the optimal distribution $q(\theta_t)$ can be shown to have the form

$$q(\theta_t) = \frac{\exp\left(\langle \log p(X,\ \Theta|\phi)\rangle_{\theta_v \neq \theta_t}\right)}{\int \exp\left(\langle \log p(X,\ \Theta|\phi)\rangle_{\theta_v \neq \theta_t}\right)\mathrm{d}\theta_t} \tag{9}$$

As will be shown, the optimal distributions $q(.)$ for the latent variables S_i, $\mu_{(l)}$, and $\Omega_{(l)}$ have the same functional forms as their respective priors $p(S_i|\phi)$, $p(\mu_{(l)})$, and $p(\Omega_{(l)})$. Specifically,

$q(S_i) = \mathrm{M}ultinomial(1, \alpha_i)$, with parameter $\alpha_i = (\alpha_{i1}, \alpha_{i2}, .., \alpha_{iK})$ where $\sum_{l=1}^{K} \alpha_{il} = 1$,

$q(\mu_{(l)}) = \mathbb{N}_d(m_{(l)},\ T_{(l)})$, a multivariate Gaussian with d-dimensional mean vector $m_{(l)}$ and $d \times d$ precision matrix $T_{(l)}$,

$q(\Omega_{(l)}) = \mathbb{W}(\eta_{(l)},\ C_{(l)})$, a Wishart distribution with degrees of freedom $\eta_{(l)}$ and $d \times d$ scale matrix $C_{(l)}$.

For the $\Lambda_{(l):i}$ latent variable vectors, the optimal form of $q(\Lambda_{(l):i})$ is quite unwieldy to work with; thus, instead, we define each $q(\Lambda_{(l):i})$ to be a multivariate Gaussian distribution with a *diagonal* precision matrix. Specifically,

$$q(\Lambda_{(l):i}) = \mathbb{N}_d(\delta_{(l)i},\ D_{(l)i}) = \prod_{j=1}^{d} q(\lambda_{lji}) = \prod_{j=1}^{d} \mathbb{N}\left(a_{lji},\ \frac{1}{b_{lji}}\right)$$

with d-dimensional mean vector $\delta_{(l)i}$ and $d \times d$ precision matrix $D_{(l)i}$.

Since $D_{(l)i}$ is a diagonal matrix, the multivariate distribution $q(\Lambda_{(l):i})$ can be written as a product of d independent *univariate* Gaussian distributions $q(\lambda_{lji})$, $1 \leq j \leq d$, where $\delta_{(l)i} = (a_{l1i}, a_{l2i}, .., a_{ldi})^T$ and $\text{diag}(D_{(l)i}) = \left(\frac{1}{b_{l1i}}, \frac{1}{b_{l2i}}, .., \frac{1}{b_{ldi}}\right)$. That is, a_{lji} and b_{lji} denote the mean and variance, respectively, of the random variable λ_{lji}.

We use Equations (8) and (9) to derive update equations for the parameters of the factor distributions. These update equations are linked, in the sense that the update equation for a variational parameter is a function of other variational parameters. Given these equations, our proposed variational EM algorithm involves maximizing the ELBO function using an iterative procedure. In each iteration, we cycle through the set of update equations to update the values of the variational parameters. While the variational parameters α_i, $m_{(l)}$, $T_{(l)}$, $\eta_{(l)}$, and $C_{(l)}$ have closed forms for their update equations (in terms of other parameters), the update values for variational parameters a_{lji} and b_{lji} are not closed form but are obtained using the Newton-Raphson method (Press et al., 1992).

Finally, the mixing coefficients are re-estimated to improve the approximation to the marginal log-likelihood. Convergence of this iterative procedure is guaranteed since the ELBO function increases with each update, unless it is already at a (local) maximum value (Boyd and Vandenberghe, 2004; Bishop, 2006). The algorithm is run on the input sample–taxa matrix using a reasonably large value of K, and after convergence, the number of mixing coefficients that are above a preset threshold value denotes the optimal number of components in the model.

2.5. *Parameter update formulas and marginal log-likelihood lower bound using ELBO function*

The update equations for the factor distribution parameters α_i, $m_{(l)}$, $T_{(l)}$, $\eta_{(l)}$, and $C_{(l)}$ are given in Table 1. The derivations of the update equations are provided in the Appendix A1. The estimates of the parameters for each $q(\Lambda_{(l):i})$ are obtained by maximizing the ELBO function restricted to each $\Lambda_{(l):i}$.

Table 1. The Update Formulas for the Parameters α_i, $m_{(l)}$, $T_{(l)}$, $\eta_{(l)}$, and $C_{(l)}$

Parameter α_i for $q(S_i)$:

$$\alpha_{il} = \frac{f_{il}}{\sum_{r=1}^{K} f_{ir}}, \text{ where}$$

$$f_{il} = \exp\{\log \phi_l - \frac{d}{2}\log(2\pi) + \frac{1}{2}\langle \log|\Omega_{(l)}|\rangle + \sum_{j=1}^{d}\left[-\langle e^{\lambda_{lji}}\rangle - \log(x_{ji}!) + \langle \lambda_{lji}\rangle x_{ji}\right]$$

$$- \frac{1}{2}\text{Tr}\left(\langle\Omega_{(l)}\rangle\left[\langle\Lambda_{(l):i}\Lambda_{(l)i}^T\rangle - \langle\mu_{(l)}\rangle\langle\Lambda_{(l):i}\rangle^T - \langle\Lambda_{(l):i}\rangle\langle\mu_{(l)}\rangle^T + \langle\mu_{(l)}\mu_{(l)}^T\rangle\right]\right)\}$$

Parameters $m_{(l)}$ and $T_{(l)}$ for $q(\mu_{(l)})$:

$$T_{(l)} = \beta I + \langle\Omega_{(l)}\rangle\sum_{i=1}^{n}\langle s_{il}\rangle,\ m_{(l)} = T_{(l)}^{-1}\langle\Omega_{(l)}\rangle\sum_{i=1}^{n}\langle\Lambda_{(l):i}\rangle\langle s_{il}\rangle$$

Parameters $\eta_{(l)}$ and $C_{(l)}$ for $q(\Omega_{(l)})$:

$$\eta_{(l)} = \nu + \sum_{i=1}^{n}\langle s_{il}\rangle$$

$$C_{(l)} = V + \sum_{i=1}^{n}\langle\Lambda_{(l):i}\Lambda_{(l):i}^T\rangle - \sum_{i=1}^{n}\langle\Lambda_{(l):i}\rangle\langle s_{il}\rangle\langle\mu_{(l)}\rangle^T - \langle\mu_{(l)}\rangle\sum_{i=1}^{n}\langle\Lambda_{(l):i}\rangle^T\langle s_{il}\rangle + \langle\mu_{(l)}\mu_{(l)}^T\rangle\sum_{i=1}^{n}\langle s_{il}\rangle$$

Expected values used in the above updates:

$$\langle s_{il}\rangle = \alpha_{il}, \qquad \langle\Lambda_{(l):i}\Lambda_{(l):i}^T\rangle = D_{(l)}^{-1} + \delta_{(l)i}\delta_{(l)i}^T$$

$$\langle\mu_{(l)}\rangle = m_{(l)}, \qquad \langle\mu_{(l)}\mu_{(l)}^T\rangle = T_{(l)}^{-1} + m_{(l)}m_{(l)}^T$$

$$\langle\lambda_{lji}\rangle = a_{lji}, \qquad \langle\Omega_{(l)}\rangle = \eta_{(l)}C_{(l)}^{-1}$$

$$\langle e^{\lambda_{lji}}\rangle = e^{a_{lji}+\frac{1}{2}b_{lji}}, \quad \langle\log|\Omega_{(l)}|\rangle = d\log 2 - \log|C_{(l)}| + \sum_{j=1}^{d}\psi\left(\frac{\eta_{(l)} + 1 - j}{2}\right)$$

$$\langle\Lambda_{(l):i}\rangle = \delta_{(l)i}$$

Notation: Tr(.) and $\psi(.)$ denote the matrix trace function and the di-gamma function, respectively; $\langle F(\theta_t)\rangle$ denotes the expectation of function $F(\theta_t)$ over the factor distribution $q(\theta_t)$

The ELBO function given in Equation (8) when restricted to $\Lambda_{(l):i}$ has the form

$$\int q(\Lambda_{(l):i}) \log p(X_{:i},\ \Lambda_{(l):i} | \langle \mu_{(l)} \rangle,\ \langle \Omega_{(l)} \rangle)^{\langle s_{il} \rangle} \mathrm{d}\Lambda_{(l):i} \ -\ \int q(\Lambda_{(l):i}) \log q(\Lambda_{(l):i}) \mathrm{d}\Lambda_{(l):i} \ +\ \text{constant}$$

The above integral can be rewritten as

$$\langle s_{il} \rangle \int \prod_{j=1}^{d} \mathbb{N}(\lambda_{lji} | a_{lji},\ 1/b_{lji}) \left[\sum_{j=1}^{d} \left[-e^{\lambda_{lji}} \ +\ \lambda_{lji} x_{ji} \right] - \frac{1}{2} \Lambda_{(l):i}^{T} \langle \Omega_{(l)} \rangle \Lambda_{(l):i} \ +\ \Lambda_{(l):i}^{T} \langle \Omega_{(l)} \rangle \langle \mu_{(l)} \rangle \right] \mathrm{d}\lambda_{l1i} \mathrm{d}\lambda_{l2i} .. \mathrm{d}\lambda_{ljI}$$
$$+\ \sum_{j=1}^{d} \log b_{lji} \ +\ \text{constant}$$

Let ω_{lrt} denote the entry in row r and column t of matrix $\langle \Omega_{(l)} \rangle$, and h_{lt} denote the t^{th} entry in vector $\langle \mu_{(l)} \rangle$. Since λ_{lji} has a univariate Gaussian distribution, it follows that $\langle \lambda_{lji} \rangle = a_{lji}$, $\langle \lambda_{lji}^2 \rangle = a_{lji}^2 + b_{lji}$, and $\langle e^{\lambda_{lji}} \rangle = e^{a_{lji} + \frac{1}{2} b_{lji}}$. We use these observations to simplify the above expression to produce a function $F(a_{l1i}, b_{l1i}, a_{l2i}, b_{l2i}, .., a_{ldi}, b_{ldi})$ of $2d$ variables. We identify the values of the variables $a_{l1i}, b_{l1i}, a_{l2i}, b_{l2i}, .., a_{ldi}, b_{ldi}$ that maximize $F(.)$. This is accomplished by cycling through each of the $2d$ variables and maximizing the corresponding univariate function on that variable.

Restricted to variables a_{lji} and b_{lji} (and excluding constants), the function $F(.)$ reduces to

$$G(a_{lij},\ b_{lij}) = \langle s_{il} \rangle \left[\sum_{j=1}^{d} [-e^{a_{lji} + \frac{1}{2} b_{lji}} \ +\ x_{ji} a_{lji}] \ -\ \frac{1}{2} \sum_{j=1}^{d} \omega_{ljj} [a_{lji}^2 \ +\ b_{lji}] \right.$$
$$\left. -\ \left[\sum_{\substack{t=1 \\ t \neq j}}^{d} \omega_{ljt}\ a_{lti} \right] a_{lji} \ +\ \left[\sum_{t=1}^{d} \omega_{ljt} h_t \right] a_{lji} \right]$$
$$+\ \frac{1}{2} \sum_{j=1}^{d} \log b_{lji}$$

As part of maximizing $F(.)$, we first compute the derivatives of $G(.)$ with respect to a_{lji} and b_{lji} separately, and set the two resulting derivatives to 0. The roots of these two equations are then computed using the Newton-Raphson method. The corresponding equations for a_{lji} and b_{lji} are, respectively,

$$H_1(a_{lji}) = -e^{(a_{lji} + \frac{1}{2} b_{lji})} \ -\ \omega_{ljj} a_{lji} \ +\ x_{ji} \ +\ \sum_{t=1}^{d} \omega_{ljt} h_{lt} \ -\ \sum_{\substack{t=1 \\ t \neq j}}^{d} \omega_{ljt} a_{lti} = 0$$

$$H_2(b_{lji}) = e^{(\frac{1}{2} b_{lji} + a_{lji})} \ -\ \frac{1}{\langle s_{il} \rangle b_{lji}} \ +\ \omega_{ljj} = 0$$

Once we have estimates for the variational parameters, we can compute the ELBO function using an expansion of Equation (8) as

$$ELBO(Q) = \langle \log p(X,\ \mathcal{L} | \mathcal{S},\ \mathcal{M},\ \mathcal{T}) \rangle + \langle \log p(\mathcal{S} | \phi) \rangle + \langle \log p(\mathcal{M}) \rangle + \langle \log p(\mathcal{T}) \rangle$$
$$- \langle \log q(\mathcal{S}) \rangle - \langle \log q(\mathcal{M}) \rangle - \langle \log q(\mathcal{T}) \rangle - \langle \log q(\mathcal{L}) \rangle$$

The formulas for the expected values in Equation (10) are given in Table 2. Since the ELBO function approximates the true marginal log-likelihood function $\log p(X|\phi)$, after we have cycled through and estimated the variational parameters, we can then maximize the resulting ELBO with respect to the mixing coefficients. This can be done by taking the derivatives of Equation (10) with respect to the ϕ_l's and using a Lagrange multiplier to enforce the constraint that $\sum_{l=1}^{K} \phi_l = 1$. It can be shown that $\phi_l = \frac{1}{n} \sum_{i=1}^{n} \alpha_{il}$, for $1 \leq l \leq K$ (Bishop, 2006).

2.6. Variational EM algorithm (MS_MixMPLN)

Input: Sample–taxa matrix $X_{d \times n}$, number of components K, the prior parameters β, ν, and V.

Output: Values of the mixing coefficients and the variational parameters that maximize the ELBO function, and the maximum ELBO function value.

Table 2. The Equations to Compute the Expected Values of the Different Constituents of the Evidence Lower Bound Function

$$\langle \log p(X, \mathcal{L}|\mathcal{S}, \mathcal{M}, \mathcal{T})\rangle = \sum_{l=1}^{K}\sum_{i=1}^{n}\langle s_{il}\rangle \left[-\frac{d}{2}\log(2\pi) + \frac{1}{2}\langle \log|\Omega_{(l)}|\rangle + \sum_{j=1}^{d}\left[-\langle e^{\lambda_{lji}}\rangle - \log(x_{ji}!) + \langle \lambda_{lji}\rangle x_{ji}\right] - \frac{1}{2}\mathrm{Tr}\left(\langle\Omega_{(l)}\rangle\left[\langle\Lambda_{(l):i}\Lambda_{(l):i}^{T}\rangle - \langle\mu_{(l)}\rangle\langle\Lambda_{(l):i}\rangle^{T} - \langle\Lambda_{(l):i}\rangle\langle\mu_{(l)}\rangle^{T} + \langle\mu_{(l)}\mu_{(l)}^{T}\rangle\right]\right)\right]$$

$$\langle \log p(\mathcal{S}|\phi)\rangle = \sum_{l=1}^{K}\sum_{i=1}^{n}\langle s_{il}\rangle \log\phi_l$$

$$\langle \log p(\mathcal{M})\rangle = \frac{Kd}{2}\log\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2}\sum_{l=1}^{K} Tr(\langle\mu_{(l)}\mu_{(l)}^{T}\rangle)$$

$$\langle \log p(\mathcal{T})\rangle = K\left[-\frac{\nu d}{2}\log 2 - \frac{d[d-1]}{4}\log\pi - \sum_{j=1}^{d}\log\Gamma\left(\frac{\nu+1-j}{2}\right) + \frac{\nu}{2}\log|V|\right] + \frac{\nu-d-1}{2}\left[\sum_{l=1}^{K}\langle\log|\Omega_{(l)}|\rangle\right] - \frac{1}{2}Tr\left(V\sum_{l=1}^{K}\langle\Omega_{(l)}\rangle\right)$$

$$\langle \log q(\mathcal{S})\rangle = \sum_{i=1}^{n}\sum_{l=1}^{K}\langle s_{il}\rangle\log\langle s_{il}\rangle$$

$$\langle \log q(\mathcal{M})\rangle = \sum_{l=1}^{K}\langle\log q(\mu_{(l)})\rangle = \sum_{l=1}^{K}\left[-\frac{d}{2}[1+\log 2\pi] + \frac{1}{2}\log|T_{(l)}|\right]$$

$$\langle \log q(\mathcal{T})\rangle = \sum_{l=1}^{K}\langle\log q(\Omega_{(l)})\rangle = \sum_{l=1}^{K}\left[-\frac{d\eta_{(l)}}{2}\log 2 - \frac{d[d-1]}{4}\log\pi - \sum_{j=1}^{d}\log\Gamma\left(\frac{\eta_{(l)}+1-j}{2}\right) + \frac{\eta_{(l)}}{2}\log|C_{(l)}| + \frac{\eta_{(l)}-d-1}{2}\langle\log|\Omega_{(l)}|\rangle - \frac{1}{2}Tr(C_{(l)}\langle\Omega_{(l)}\rangle)\right]$$

$$\langle \log q(\mathcal{L})\rangle = \sum_{i=1}^{n}\sum_{l=1}^{K}\langle\log q(\Lambda_{(l):i})\rangle = \sum_{i=1}^{n}\sum_{l=1}^{K}\left[\frac{1}{2}\left[\sum_{j=1}^{d}\log b_{lji}\right] - \frac{d}{2}[1+\log 2\pi]\right]$$

Notation: $\Gamma(.)$ denotes the gamma function.

Initialization: Initialize the mixing coefficient vector $\phi = (\phi_1, \phi_2, .., \phi_K)$ and the variational parameters α_i, $m_{(l)}$, $T_{(l)}$, $\eta_{(l)}$, $C_{(l)}$, $\delta_{(l)i}$, and $D_{(l)i}$, for $1 \le i \le n$, $1 \le l \le K$.

Repeat until convergence (i.e., the ELBO function does not increase any further):

E-step:

Cycle through the variational parameters and update their estimates.

M-step:

Set $\phi_l = \frac{1}{n}\sum_{i=1}^{n}\alpha_{il}$, for $1 \le l \le K$.

We implemented MS_MixMPLN in the R programming language (R Development Core Team, 2013). The program is available at https://github.com/syooseph/YoosephLab/tree/master/MixtureMicrobialNetworks/MS_MixMPLN

3. RESULTS

The performance of the model selection algorithm MS_MixMPLN was evaluated using a collection of synthetic sample–taxa count matrices with d taxa and n samples. The samples (count vectors) in each sample–taxa matrix were generated from a mixture model consisting of K MPLN component distributions and with mixing coefficient vector ϕ. The precision matrices of the component distributions were generated from an underlying graph structure. Five different types of graph structures were considered and these were band, cluster, hub, random, and scale free. The R *huge* package (Zhao et al., 2012) was used to generate the precision matrices associated with each graph structure. Sample–taxa matrices were generated with number of taxa $d = 50$, number of components $K = 2, 3, 4$, and number of samples $n = sK$, where s is the number of samples per component ($s = 200, 1000$). The mixing coefficient vectors for $K = 2, 3$ and 4 were $\phi = \left(\frac{1}{2}, \frac{1}{2}\right)$, $\phi = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$, and $\phi = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$, respectively. For each graph type and combination of parameter values, 20 replicates were generated. Thus, 600 synthetic sample–taxa matrices were generated in total.

For each input sample–taxa matrix, MS_MixMPLN was run with a larger value for the number of components (5, 6, and 7, respectively, for ground-truth $K=2$, 3, and 4). The values for the priors were as follows: $\beta=10^{-6}$, $\nu=50$, and V set to a diagonal matrix with all entries equal to 51. MS_MixMPLN was run using 26 different starting points on each input, where 25 starting points were generated using random partitions of the samples in the sample–taxa matrix and one starting point was generated using the K-means algorithm (MacQueen, 1967) to partition the samples. Each partition was used to initialize the estimates for the mixing coefficients and the variational parameters. For each starting point, the algorithm was allowed to run for 30 iterations (or less, if it reached convergence). The output with the maximum ELBO value was selected.

The predicted number of components was determined by applying a threshold τ to the mixing coefficient values; that is, any component with mixing coefficient value $< \tau$ was not counted toward the predicted number of components. Table 3 shows the accuracy of MS_MixMPLN on the simulated datasets at different values for the threshold τ. Here, *accuracy* for a particular τ value is measured as the proportion of times the predicted number of components at that threshold is equal to the ground-truth K.

We see from Table 3 that, for each combination of n, K, and τ values, MS_MixMPLN shows fairly similar accuracy levels for all graph structures. While the accuracy is highest for $K=2$, in the region of 0.75 to 0.95 (at $\tau=0.01$), it decreases for $K=3$ and 4, to smaller values in the region of 0.05 to 0.25 (at $\tau=0.01$). The accuracy also generally increases with an increase in the number of samples per component (from 200 to 1000). At a higher threshold value ($\tau=0.06$), even for $K=4$, the accuracy estimates are moderately high (in the region of 0.5 to 0.75) for the sample sizes explored.

Table 3. Accuracy of MS_MixMPLN on the Simulated Dataset for Different Values of Threshold τ

Graph type	*No. of components*	*No. of samples per component*	*Threshold* $\tau=0.01$	*Threshold* $\tau=0.02$	*Threshold* $\tau=0.03$	*Threshold* $\tau=0.04$	*Threshold* $\tau=0.05$	*Threshold* $\tau=0.06$
Band	$K=2$	200	0.95	0.95	0.95	0.95	0.95	0.95
		1000	0.65	0.75	0.75	0.75	0.75	0.8
	$K=3$	200	0.15	0.3	0.4	0.45	0.5	0.55
		1000	0.2	0.25	0.3	0.5	0.55	0.6
	$K=4$	200	0.05	0.2	0.35	0.45	0.55	0.6
		1000	0.05	0.05	0.3	0.5	0.5	0.55
Cluster	$K=2$	200	0.8	0.9	0.9	0.95	0.95	0.95
		1000	0.85	0.9	0.95	0.95	1	1
	$K=3$	200	0.4	0.5	0.5	0.55	0.65	0.7
		1000	0.35	0.45	0.55	0.6	0.7	0.7
	$K=4$	200	0.2	0.25	0.25	0.25	0.35	0.4
		1000	0.05	0.05	0.15	0.25	0.4	0.6
Hub	$K=2$	200	0.8	0.9	0.9	0.9	0.9	0.85
		1000	0.75	0.9	0.95	0.95	0.95	0.95
	$K=3$	200	0.55	0.5	0.55	0.6	0.6	0.65
		1000	0.45	0.6	0.65	0.7	0.7	0.75
	$K=4$	200	0.3	0.25	0.25	0.35	0.35	0.45
		1000	0.4	0.45	0.45	0.6	0.75	0.75
Random	$K=2$	200	0.85	0.9	0.9	0.9	0.9	0.95
		1000	0.65	0.65	0.7	0.75	0.75	0.8
	$K=3$	200	0.25	0.25	0.35	0.4	0.5	0.5
		1000	0.25	0.4	0.45	0.5	0.6	0.6
	$K=4$	200	0.15	0.15	0.2	0.2	0.3	0.4
		1000	0.15	0.2	0.35	0.4	0.4	0.5
Scale-free	$K=2$	200	0.75	0.75	0.75	0.75	0.85	0.85
		1000	0.65	0.8	0.8	0.85	0.85	0.9
	$K=3$	200	0.3	0.35	0.4	0.45	0.55	0.6
		1000	0.4	0.45	0.5	0.5	0.5	0.6
	$K=4$	200	0.05	0.15	0.3	0.4	0.45	0.5
		1000	0.15	0.25	0.4	0.45	0.55	0.6

MPLN, multivariate Poisson log-normal.

As an alternate strategy, we evaluated the AIC, BIC, and EBIC criteria (Epskamp and Fried, 2018; Zhu and Cribben, 2018) for model selection. As part of this assessment, we ran the MixMPLN algorithm (Tavakoli and Yooseph, 2019) with different values for K (from 1 through 6) on each sample–taxa matrix, and used the minimum values of Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Extended BIC (EBIC) (Appendix A1) to predict the number of components. We observed that none of the three criteria returned the correct answer on any of the input datasets.

4. DISCUSSION

In this study, we presented a variational EM algorithm for selecting the number of component distributions for the MixMPLN framework. The proposed algorithm was evaluated using a large simulated dataset. For the sample sizes evaluated, the prediction accuracy decreased as the number of components increased. Future work will explore further the relationship between the number of components and the number of samples in the context of improving approximation of the marginal log-likelihood estimate and the accuracy of the algorithm. It will also include a more comprehensive examination of the parameter space and their effect on model selection accuracy. This will include exploring additional values for the prior parameters and the mixing coefficients. Finally, we will explore the ELBO function landscape further, including the use of local maxima solutions to inform the model selection process. We will also evaluate the proximity of the variational approximation to the true posterior distribution (Yao et al., 2018; Huggins et al., 2020).

AUTHORS' CONTRIBUTIONS

S.Y. and S.T. designed the algorithm. S.Y. implemented the algorithm and performed the evaluations. S.Y. wrote the article, with input from S.T.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no conflicting financial interests.

FUNDING INFORMATION

This material is based upon work supported by the National Science Foundation (NSF) under grant number DBI-2051283 to S.Y. Publication costs for this work were funded by NSF grant number DBI-2051283.

REFERENCES

Aitchison, J., and Ho, C.H. 1989. The multivariate poisson-log normal distribution. *Biometrika* 76, 643–653.

Bishop, C.M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, 2006. ISBN 0387310738.

Biswas, S., McDonald, M., Lundberg, D.S., et al. 2016. Learning microbial interaction networks from metagenomic count data. *J. Comput. Biol*. 23, 526–535.

Boyd, S., and Vandenberghe, L. 2004. *Convex Optimization*. Cambridge University Press, New York, NY, USA.

Chiquet, J., Robin, S., and Mariadassou, M. 2019. Variational inference for sparse network reconstruction from count data. *In* Kamalika, C. and Ruslan, S., eds. *Proceedings of the 36th International Conference on Machine Learning*, volume 97. PMLR, Long Beach, CA, USA.

Corduneanu, A., and Bishop, C.M. 2001. Variational bayesian model selection for mixture distributions. *In* Jaakkola, T. and Richards, T., eds. *Proceedings Eighth International Conference on Artificial Intelligence and Statistics*. Morgan Kaufmann. pp. 27–34.

Epskamp, S., and Fried, E.I. 2018. A tutorial on regularized partial correlation networks. *Psychol. Methods* 23, 617–634.

Falkowski, P.G., Fenchel, T., and Delong, E.F. 2008. The microbial engines that drive earth's biogeochemical cycles. *Science* 320, 1034–1039.

Fang, H., Huang, C., Zhao, H., et al. 2015. Cclasso: Correlation inference for compositional data through lasso. *Bioinformatics* 31, 3172–3180.

Friedman, J., and Alm, E.J. 2012. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol*. 8, e1002687.

Hibbing, M.E., Fuqua, C., Parsek, M.R., et al. 2010. Bacterial competition: Surviving and thriving in the microbial jungle. *Nat. Rev. Microbiol*. 8, 15–25.

Holmes, I., Harris, K., and Quince, C. 2012. Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PLoS One* 7, e30126.

Huggins, J., Kasprzak, M., Campbell, T., et al. 2020. Validated variational inference via practical posterior error bounds. *In* Silvia, C. and Roberto, C., eds. *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108. PMLR.

Inouye, D.I., Yang, E., Allen, G.I., et al. 2017. A review of multivariate distributions for count data derived from the poisson distribution. *WIREs Comput. Stat.* 9, n/a.

Kurtz, Z.D., Muller, C.L., Miraldi, E.R., et al. 2015. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol*. 11, e1004226.

La Rosa, P.S., Brooks, J.P., Deych, E., et al. 2012. Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLoS One* 7, e52078.

Lange, K. 2016. *MM Optimization Algorithms*. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics. [Epub ahead of print]; DOI: 10.1137/1.9781611974409.

Layeghifard, M., Hwang, D.M., and Guttman, D.S. 2017. Disentangling interactions in the microbiome: A network perspective. *Trends Microbiol*. 25, 217–228.

Loftus, M., Hassouneh, S.A., and Yooseph, S. 2021. Bacterial associations in the healthy human gut microbiome across populations. *Sci. Rep*. 11, 2828.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. *In* Le, L.M. and Neyman, J., eds. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, *Fifth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley. pp. 281–297.

Methe, B.A., Nelson, K.E., Pop, M., et al. 2012. A framework for human microbiome research. *Nature* 486, 215–221.

Parisi, G. 1988. *Statistical Field Theory*. Addison-Wesley, Redwood City.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., et al. 1992. *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. Cambridge University Press, Cambridge [Cambridgeshire]; New York.

Quail, M.A., Smith, M., Coupland, P., et al. 2012. A tale of three next generation sequencing platforms: Comparison of ion torrent, pacific biosciences and illumina miseq sequencers. *BMC Genomics* 13, 341.

R Development Core Team. 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rusch, D.B., Halpern, A.L., Sutton, G., et al. 2007. The sorcerer ii global ocean sampling expedition: Northwest atlantic through eastern tropical pacific. *PLoS Biol* 5, e77.

Tavakoli, S., and Yooseph, S. 2019. Learning a mixture of microbial networks using minorization-maximization. *Bioinformatics* 35, i23–i30.

Tyson, G.W., Chapman, J., Hugenholtz, P., et al. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428, 37–43.

Tzikas, D.G., Likas, A.C., and Galatsanos, N.P. 2008. The variational approximation for bayesian inference. *IEEE Signal Process. Mag.* 25, 131–146.

Venter, J.C., Remington, K., Heidelberg, J.F., et al. 2004. Environmental genome shotgun sequencing of the sargasso sea. *Science* 304, 66–74.

Vogel, T.M., Simonet, P., Jansson, J.K., et al. 2009. Terragenome: A consortium for the sequencing of a soil metagenome. *Nat. Rev. Microbiol*. 7, 252–252.

Williamson, S.J., and Yooseph, S. 2012. From bacterial to microbial ecosystems (metagenomics). *Methods Mol Biol*. 804, 35–55.

Wishart, J. 1928. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika* 20A, 32–52.

Woese, C.R., and Fox, G.E. 1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci. USA*. 74, 5088–5090.

Yao, Y., Vehtari, A., Simpson, D., et al. 2018. Yes, but did it work?: Evaluating variational inference. *In* Jennifer, D. and Andreas, K., eds. *Proceedings of the 35th International Conference on Machine Learning*, volume 80. PMLR.

Zhao, T., Liu, H., Roeder, K., et al. 2012. The huge package for high-dimensional undirected graph estimation in r. *J. Mach. Learn. Res*. 13:1059–1062.

Zhu, Y., and Cribben, I. 2018. Sparse graphical models for functional connectivity networks: Best methods and the autocorrelation issue. *Brain Connect*. 8, 139–165.

Address correspondence to:
Prof. Shibu Yooseph
Department of Computer Science
Genomics and Bioinformatics Cluster
University of Central Florida
Orlando, FL 32816
USA

E-mail: shibu.yooseph@ucf.edu

Appendix

APPENDIX A1

UPDATE EQUATIONS FOR THE VARIATIONAL PARAMETERS

1. For $q(S_i)$:

The optimal form for $q(S_i) = \frac{\exp(\langle \log p(X, \Theta|\phi)\rangle_{\theta \neq S_i})}{\sum_{S_i} \exp(\langle \log p(X, \Theta|\phi)\rangle_{\theta \neq S_i})}$, where $\theta \in \Theta$ and the summation in the denominator is overall possible 1-of-K binary vectors.

This quantity can be shown to be equal to

$$\frac{\exp\left(\sum_{l=1}^{K} \log (f_{il})^{s_{il}}\right)}{\sum\limits_{S_i} \exp\left(\sum_{l=1}^{K} \log (f_{il})^{s_{il}}\right)} = \frac{\prod_{l=1}^{K} [f_{il}]^{s_{il}}}{\sum_{r=1}^{K} f_{ir}} = \prod_{l=1}^{K} \alpha_{il}^{s_{il}}$$

where $\alpha_{il} = \frac{f_{il}}{\sum_{r=1}^{K} f_{ir}}$, and

$$f_{il} = \exp\Bigg\{ \log \phi_l - \tfrac{d}{2} \log (2\pi) + \tfrac{1}{2} \langle \log |\Omega_{(l)}| \rangle + \sum_{j=1}^{d} \left[-\langle e^{\lambda_{lji}} \rangle - \log (x_{ji}!) + \langle \lambda_{lji} \rangle x_{ji} \right] - \tfrac{1}{2} \mathrm{Tr} \Big(\langle \Omega_{(l)} \rangle \left[\langle \Lambda_{(l):i} \Lambda_{(l):i}^T \rangle - \langle \mu_{(l)} \rangle \langle \Lambda_{(l):i} \rangle^T - \langle \Lambda_{(l):i} \rangle \langle \mu_{(l)} \rangle^T + \langle \mu_{(l)} \mu_{(l)}^T \rangle \right] \Big) \Bigg\}$$

2. For $q(\mu_l)$:

We first observe that for a multivariate Gaussian distribution $\mathbb{N}(y|m, P)$, where m and P are the mean vector and precision matrix, respectively,

$$\log (\mathbb{N}(y|m, P)) = -\frac{1}{2} y^T P y + y^T P m + \text{constant} \tag{11}$$

where the constant is independent of y.

The form of the optimal distribution for $q(\mu_{(l)}) = \frac{\exp(\langle \log p(X,\ \Theta|\phi)\rangle_{\theta\neq\mu_{(l)}})}{\int \exp(\langle \log p(X,\ \Theta|\phi)\rangle_{\theta\neq\mu_{(l)}})\, d\mu_{(l)}}$.

The above equation can be expanded and the numerator can be shown to be equal to

$$-\frac{1}{2}[\mu_{(l)}^T\langle\Omega_{(l)}\rangle\sum_{i=1}^{n}\langle s_{il}\rangle\mu_{(l)} + \mu_{(l)}^T\beta I\mu_{(l)}] + \mu_{(l)}^T\sum_{i=1}^{n}\langle\Lambda_{(l):i}\rangle\langle s_{il}\rangle + \text{constant}$$

where the constant term is independent of $\mu_{(l)}$.

Comparing this quantity to Equation (11), we can deduce that the optimal $q(\mu_{(l)})$ is a multivariate Gaussian distribution with mean vector $m_{(l)}$ and precision matrix $T_{(l)}$ such that

$$m_{(l)} = T_{(l)}^{-1}\langle\Omega_{(l)}\rangle\sum_{i=1}^{n}\langle\Lambda_{(l):i}\rangle\langle s_{il}\rangle$$
$$T_{(l)} = \beta I + \langle\Omega_{(l)}\rangle\sum_{i=1}^{n}\langle s_{il}\rangle$$

Thus, $\langle\mu_{(l)}\rangle = m_{(l)}$. Also, since $q(\mu_l)$ is a multivariate Gaussian distribution, we have that $\langle\mu_{(l)}\mu_{(l)}^T\rangle = T_{(l)}^{-1} + m_{(l)}m_{(l)}^T$.

3. For $q(\Omega_{(l)})$:

A similar approach can be used to show that $q(\Omega_{(l)})$ is a Wishart distribution. We note that for a Wishart distribution $\mathbb{W}(Y_{d\times d}|v,\ V_{d\times d})$

$$\log(\mathbb{W}(Y|v,\ V)) = \frac{v-d-1}{2}\log|Y| - \frac{1}{2}Tr(VY) + \text{constant}$$

where the constant term is independent of matrix Y.

The numerator of the optimal form for $q(\Omega_{(l)})$ is $\exp(\langle \log p(X, \Theta|\phi)\rangle_{\theta\neq\Omega_{(l)}})$ and this can be shown to be equal to

$$\begin{aligned}\exp(\langle \log p(X,\ \Theta|\phi)\rangle_{\theta\neq\Omega_{(l)}}) = {} & \frac{1}{2}\sum_{i=1}^{n}\langle s_{il}\rangle\log|\Omega_{(l)}| - \frac{1}{2}Tr([\sum_{i=1}^{n}\langle\Lambda_{(l):i}\Lambda_{(l):i}^T\rangle - \sum_{i=1}^{n}\langle\Lambda_{(l):i}\rangle\langle s_{il}\rangle\langle\mu_{(l)}\rangle^T \\ & - \langle\mu_{(l)}\rangle\sum_{i=1}^{n}\langle\Lambda_{(l):i}\rangle^T\langle s_{il}\rangle + \langle\mu_{(l)}\mu_{(l)}^T\rangle\sum_{i=1}^{n}\langle s_{il}\rangle]\Omega_{(l)}) \\ & + \frac{v-d-1}{2}\log|\Omega_{(l)}| - \frac{1}{2}Tr(V\Omega_{(l)}) + \text{constant}\end{aligned}$$

where the constant is independent of $\Omega_{(l)}$.

From the above equation, we can deduce that $q(\Omega_{(l)})$ is a Wishart distribution with degrees of freedom $\eta_{(l)}$ and scale matrix $C_{(l)}$ defined as

$$\eta_{(l)} = \nu + \sum_{i=1}^{n}\langle s_{il}\rangle$$
$$C_{(l)} = V + \sum_{i=1}^{n}\langle\Lambda_{(l):i}\Lambda_{(l):i}^T\rangle - \sum_{i=1}^{n}\langle\Lambda_{(l):i}\rangle\langle s_{il}\rangle\langle\mu_{(l)}\rangle^T - \langle\mu_{(l)}\rangle\sum_{i=1}^{n}\langle\Lambda_{(l):i}\rangle^T\langle s_{il}\rangle + \langle\mu_{(l)}\mu_{(l)}^T\rangle\sum_{i=1}^{n}\langle s_{il}\rangle$$

Thus, $\langle\Omega_{(l)}\rangle = \eta_{(l)}C_{(l)}^{-1}$. Also, $\langle\log|\Omega_{(l)}|\rangle = d\log 2 - \log|C_{(l)}| + \sum_{j=1}^{d}\psi(\frac{\eta_{(l)}+1-j}{2})$.

MODEL SELECTION USING BIC, AIC, AND EBIC

MixMPLN was run with different values of K. The K value with minimum BIC, AIC, or EBIC score (Epskamp and Fried, 2018; Zhu and Cribben, 2018) was selected as the predicted number of components. Here,

$$\mathrm{AIC} = 2k - 2\log L$$

$$\mathrm{BIC} = k\log n - 2\log L$$

$$\mathrm{EBIC} = k\log n - 2\log L + 4\gamma k\log(Kd)$$

where $\log L$ is the log-likelihood score, K is the number of components, k is the total number of non-zero elements in the precision matrices of the K components, d is the number of taxa, γ is a constant (set to 0.5), and n is the number of samples.