A Graphical Model for Fusing Diverse Microbiome Data

M. Aktukmak, H. Zhu, M. Chevrette, J. Nepper, S. Magesh, J. Handelsman, A. Hero

A Graphical Model for Fusing Diverse Microbiome Data

Mehmet Aktukmak[1,*], Haonan Zhu[1], Marc G. Chevrette[2,3], Julia Nepper[3], Shruthi Magesh[3,4], Jo Handelsman[2,3], and Alfred Hero[1,*]

[1]Department of Electrical and Computer Engineering, University of Michigan
[2]Department of Plant Pathology, University of Wisconsin-Madison
[3]Wisconsin Institute for Discovery
[4]Microbiology Doctoral Training Program, University of Wisconsin-Madison
[*]Corresponding Authors: aktukmak@umich.edu, hero@eecs.umich.edu

***Abstract*—This paper develops a Bayesian graphical model for fusing disparate types of count data. The motivating application is the study of bacterial communities from diverse high dimensional features, in this case transcripts, collected from different treatments. In such datasets, there are no explicit correspondences between the communities and each correspond to different factors, making data fusion challenging. We introduce a flexible multinomial-Gaussian generative model for jointly modeling such count data. This latent variable model jointly characterizes the observed data through a common multivariate Gaussian latent space that parameterizes the set of multinomial probabilities of the transcriptome counts. The covariance matrix of the latent variables induces a covariance matrix of co-dependencies between all the transcripts, effectively fusing multiple data sources. We present a computationally scalable variational Expectation-Maximization (EM) algorithm for inferring the latent variables and the parameters of the model. The inferred latent variables provide a common dimensionality reduction for visualizing the data and the inferred parameters provide a predictive posterior distribution. In addition to simulation studies that demonstrate the variational EM procedure, we apply our model to a bacterial microbiome dataset.[1]**

I. Introduction

We introduce a Bayesian graphical model for jointly modeling and fusing high dimensional count data collected from different sensors with no explicit correspondences between their feature sets. Our model is relevant to the many areas of multi-modality fusion where data is collected from diverse but incommensurate sensor modalities. Examples include multi-view learning in computer vision and automated language translation in natural language processing. However, this paper focuses on a particularly timely application: fusion of microbiome data from diverse microbial communities.

The analysis of meta-transcriptomic data has been of increasing interest to researchers in the biological and health sciences. Microbiomes exist in diverse environments and are critical to sustaining life, balancing ecosystems, and producing antibiotics, among many other functions. Microbiomes consist of communities of microbes that interact with each other to maintain stability and resilience to environmental conditions and microbial intrusions from competitors. It has therefore been of great scientific interest to quantify changes in microbiome communities due to changing conditions using experimental data. For example, one area of study is the rhizosphere, which is a community of microbial species living around plant root systems, known to be sensitive to environmental factors [15]. Another area of study is the spectrum of responses of microbiomes to stressors, collectively called the microbial exposome [31].

One of the principal sensing platforms used to study microbiome communities applies gene sequencing to a microbiome sample, e.g., collected from the gut, the soil, or other environment. A common way to obtain a global profile of a microbial community is to perform gene sequencing on a biological sample. For example, RNA-Seq measures gene expression in a community by quantifying the number of times each gene transcript occurs in the pool of sequenced RNAs. Each microbial species in the community is represented by its own unique set of transcripts, i.e., its transcriptome, and fusing information from different transcriptomes yields the global profile of gene expression across all species in the community. This type of analysis is known as metatranscriptomics and it provides a functional profile of the community that can complement the gene taxonomic profiling provided by metagenomics [37], [29], [1]. This paper introduces a Bayesian graphical model for the metatranscriptomics problem, where inference is performed using a scalable variational EM inference method. Notably, our model can capture patterns of similarity between histograms of different species' gene expression without inter-species genome-to-genome mappings or knowledge of inter-species transcriptomic pathway correspondences.

A main feature of our model is that it estimates the global covariance structure of gene expression when the observations are in the form of count vectors produced by RNA-Seq. Correlations between transcript abundances

[1]This work was partially supported by grants from ARO W911NF-19-102 and DOE DE-NA0003921.

are informative about the effect of environmental conditions on microbial communities [38]. In particular, the global covariance matrix captures inter- and intra species interactions. For example, the expression of a single gene in a species can influence other gene expressions in that species or the gene expressions of other community members. We propose a latent variable graphical model that can capture the hidden factors underlying such dependencies.

The main assumption underlying our proposed model is the existence of a hidden low-dimensional continuous latent space that can explain the observed data. We model the observations as conditionally multinomial distributed given the latent variables, which are assumed to be multivariate Gaussian with low rank covariance structure. Due to the lack of conjugacy between the Gaussian and multinomial distributions, exact Bayes inference is not tractable. We therefore adopt a Bayes variational inference approach [9], [6] to develop an algorithm for estimating the parameters of the proposed model and projecting the data to the latent space.

The proposed model can be contrasted with previously introduced latent variable models used in multi-view learning and dimensionality reduction. Factor analysis (FA) [34] is a classical method that is a generalization of Principal Component Analysis (PCA) [4] and Probabilistic PCA [40]. FA decomposes the observed data matrix into a low dimensional set of factor loadings and factor scores, imposing a low-rank constraint on the covariance matrix. Like our proposed model, the FA model also assumes a low-dimensional Gaussian latent space but it does not account for the counting nature of the observed data.

Several latent variable models have been proposed for counting observations. These include Latent Semantic Analysis [23], Multinomial PCA [11], and Latent Dirichlet Allocation (LDA) [7]. LDA is the most closely related model to the model proposed here since it is also a Bayesian graphical model for count data and uses multinomial distribution. The main difference is that LDA uses a Dirichlet distributed latent space instead of a Gaussian distributed latent space. Our Gaussian distributed latent space makes it possible to recover a non-trivial covariance structure among the count variables, unlike LDA [5], [33].

Another way to capture the covariance structure of the observed variables is to ignore the counting nature of the data and use Gaussian Markov random fields (GMRF) [16] to directly estimate the covariance, or Gaussian Graphical Models (GGM) [30] to enforce sparsity on the inverse of the covariance estimate. Unlike our proposed multinomial-Gaussian model the GMRF and GGM do not incorporate a low rank latent structure on the covariance nor do they account for the counting nature of the data. There have been extensions of the GGM to handle multinomial observations using copulas [26] that have been applied to microbiome analysis [35], [36], [43].

Inference in latent variable models, like the one we propose here, can be challenging. This is especially difficult when there is lack of conjugacy between the distributions of the latent variables and the observed variables. One approach is to perform point estimation for both the latent variables and the parameters in an alternating fashion [13], but this is prone to over-fitting [41] and convergence issues. Another approach is to use Markov Chain Monte Carlo (MCMC) methods, which can be computationally expensive [32], especially in high dimension. As an alternative, variational Bayes inference has shown much promise [6]. Note that Variational Bayes is not a general purpose method and must be tailored to the specific statistical model [22]. When there is a lack of conjugacy, as is the case for the multinomial-Gaussian model in this paper, local variational bound approximations are often adopted [9]. Additionally, when there is a problematic expression in the joint density, such as LogSumExp or LogGamma function, which may prevent the inference of the latent variables, surrogate optimization transfer based on Taylor series expansion can be applied to approximate the non-linear function either with linear [5] or quadratic [8], [18], [10], [20], [19] functions. We adopt such a local variational bound approach for deriving an inference algorithm for our proposed model.

We summarize our contributions as follows. First, we propose a novel multinomial-Gaussian graphical model to fuse and capture the low rank covariance structure in counting data of disparate types. Our low-dimensional continuous latent space formulation provides dimensionality reduction that can be used for visualization of the count vectors on a common space. Second, we develop a novel and computationally scalable optimization algorithm based on variational inference to fit the proposed model, which exploits variational local bound approximations. Third, we validate and illustrate the model and its inference algorithm on a synthetic dataset and a real-world bacterial microbiome dataset [2].

II. Proposed Model

In this section, we formally define our proposed model and its corresponding variational inference algorithm. Lastly, we discuss computational complexity.

A. Notation

We denote the ith data sample of the lth species as $\boldsymbol{x}_{kl,i} \in \mathbb{Z}_+^{d_l}$, where k indexes the experimental condition, and d_l denotes the total number of transcripts for species l. The total number of experimental condition from which the samples are collected is denoted as K, and the total number of species in the model community is denoted as L, hence $l = 1 : L$ and $k = 1 : K$. For each experimental condition, different numbers of identically distributed samples are collected. Hence, we denote the total number of samples for the experimental condition k as I_k. The

[2]The code and the dataset are available at https://github.com/maktukmak/microbiome-thor.

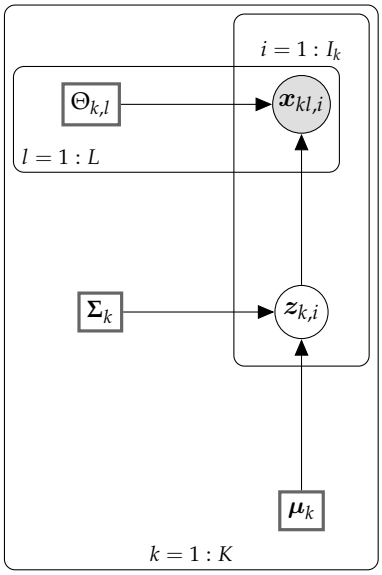


Fig. 1: Graphical model representation of the proposed latent variable model. $\boldsymbol{x}_{kl,i}$ corresponds to the ith sample of community l collected from environment k. The variables $\{\boldsymbol{x}_{kl,i}\}_{l=1:L}$ share a common low-dimensional latent variable $\boldsymbol{z}_{k,i}$ that captures the hidden causes of the observations.

count vector for experimental condition k, species l, and replicate i is denoted $\boldsymbol{x}_{kl,i}$. The dataset for experimental condition k is $D_k = \{\{\boldsymbol{x}_{kl,i}\}_{l=1}^{L}\}_{i=1}^{I_k}$.

B. Latent Variable Model

We model the observed multi-species model community data as generated from a low dimensional latent variable generative model. Under this model the data are conditionally multinomial distributed given the latent variables, which are themselves Gaussian distributed with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma_k}$. As will be shown below, the model fuses the observed data across species and it induces a low rank decomposition of the population transcriptome covariance. Let $\boldsymbol{z}_{k,i} \in \mathbb{R}^{d_z}$ be the latent variable assigned for the data sample $D_{k,i}$. $\boldsymbol{z}_{k,i}$ thus has the following multivariate normal prior distribution:

$$p(\boldsymbol{z}_{k,i}|\boldsymbol{\mu}_k, \Sigma_k) = \mathcal{N}(\boldsymbol{z}_{k,i}|\boldsymbol{\mu}_k, \Sigma_k), \tag{1}$$

where $\boldsymbol{\mu}_k \in \mathbb{R}^{d_z}$ is the prior mean vector and $\Sigma_k \in \mathbb{S}_{++}^{d_z}$ is the positive definite prior covariance matrix. The observed data consists of count vectors of the transcriptomes, which are modeled as multinomial distributed [33]. We model the conditional distributions of the observed count vectors of species l as follows:

$$p(\boldsymbol{x}_{kl,i}|\boldsymbol{z}_{k,i}, \Theta_{kl}) = Mu(\boldsymbol{x}_{kl,i}|N_{kl,i}, \mathcal{S}(\Theta_{kl}^T \boldsymbol{z}_{k,i})), \tag{2}$$

where Mu denotes the multinomial distribution, and $N_{kl,i}$ is the total number of counts of the ith data sample of species l. Note that we have introduced one more model parameter for each species, specifically $\Theta_{kl} \in \mathbb{R}^{d_l \times d_z}$, that maps lower dimensional latent space to the higher dimensional observation space of species l. Also note that both the latent variable $\boldsymbol{z}_{k,i}$ and the parameter Θ_{kl} are real-valued. Therefore, to provide a proper simplex support set for the multinomial distribution, we use the soft-max function, $\mathcal{S}(\boldsymbol{\eta})_d = \exp \eta_d / \sum_{d'=1}^{D} \exp \eta_{d'}$. The output of this function is a proper probability vector, i.e., $\sum_{d=1}^{D} \mathcal{S}(\boldsymbol{\eta})_d = 1$ and $\mathcal{S}(\boldsymbol{\eta})_d \geq 0$ for all $d = 1 : D$. See Fig. 1 for a graphical representation of the proposed model.

Although it is natural to model the observed counts as multinomial distributed, it may not be obvious why we use Gaussian latent variables for the latent space. A conjugate distribution such as Dirichlet may seem more natural than the Gaussian distribution, which is not conjugate to Multinomial. However, the components of the Dirichlet distribution are nearly independent [5], hence it is non-trivial to capture the correlations between the hidden components. On the other hand, the Multivariate normal distribution has a covariance parameter which specifically captures the correlation between the hidden components. This is useful for modeling the correlation between multiple datasets. Similar model assumptions are also adopted in topic models [5], [39], categorical PCA [20], and Gaussian process classification [19]. Note that, although the communities are dependent through the latent variables, the experimental conditions are modeled as independent. Hence, there is no coupling between the experimental conditions and thus we fit independent models for each condition.

The joint log likelihood of the proposed model is of the form $\sum_{k=1}^{K} \sum_{i=1}^{I_k} \log p(\boldsymbol{z}_{k,i}, D_{k,i})$, where:

$$\begin{aligned} \log p(\boldsymbol{z}_{k,i}, D_{k,i}) &= \log p(\boldsymbol{z}_{k,i}) + \sum_{l=1}^{L} \log p(\boldsymbol{x}_{kl,i}|\boldsymbol{z}_{k,i}) \\ &= -\frac{1}{2}\Big[(\boldsymbol{z}_{k,i} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\boldsymbol{z}_{k,i} - \boldsymbol{\mu}_k) + \log |\Sigma_k|\Big] \\ &\quad + \sum_{l=1}^{L} \sum_{d=1}^{D} \boldsymbol{x}_{kl,id} (\Theta_{kl,d} \boldsymbol{z}_{k,i} - \mathrm{lse}(\Theta_{kl} \boldsymbol{z}_{k,i})) + const, \end{aligned} \tag{3}$$

in which lse denotes the log-sum-exp function, i.e., log of the denominator of the soft-max function, and we suppress the deterministic parameters to avoid clutter. Taking the expectation with respect to $\boldsymbol{z}_{k,i}$ is tractable for the linear and quadratic terms, but intractable for the lse term. We describe an asymptotic approximation in the next section.

C. Optimization

Next, we develop a variational EM maximum likelihood algorithm [9], [6] to infer the deterministic parameters $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, and Θ_{kl}. The main objective is to maximize the likelihood of the observations under the model. The algorithm comprises two alternating steps: i) Expectation step (E-step), where we integrate out the latent variables, ii) Maximization step (M-step), where we optimize the model parameters to maximize the marginal likelihood.

1) Objective: The proposed model uses Gaussian latent variables for the multinomial observations. Due to lack of conjugacy between Gaussian and Multinomial distributions, the likelihood function is not closed form. Specifically, integrating out the latent variables becomes intractable (See Section II-C2 for the details). Hence, we resort to variational inference, in which a lower bound on the likelihood function is derived and maximized. This lower bound is obtained by approximating the posterior distributions of the latent variables. In variational inference, the objective is to minimize the distance (KL-divergence) between the approximate and exact posterior distributions. This objective can be expressed for a single latent variable $\boldsymbol{z}_{k,i}$ as follows:

$$\begin{aligned}\mathbb{KL}(q_{\lambda_{k,i}}|p) &= \mathbb{E}_{q_{\lambda_{k,i}}} \log\Big[\frac{q(\boldsymbol{z}_{k,i}|\boldsymbol{\lambda}_{k,i})}{p(\boldsymbol{z}_{k,i}|D_{k,i})}\Big] \\ &= \mathbb{E}_{q_{\lambda_{k,i}}} \log\Big[\frac{q(\boldsymbol{z}_{k,i}|\boldsymbol{\lambda}_{k,i})}{p(\boldsymbol{z}_{k,i},D_{k,i})}p(D_{k,i})\Big] \\ &= \mathbb{E}_{q_{\lambda_{k,i}}}\Big[\log q(\boldsymbol{z}_{k,i}|\boldsymbol{\lambda}_{k,i}) - \log p(\boldsymbol{z}_{k,i},D_{k,i})\Big] \\ &\quad + \log p(D_{k,i}),\end{aligned} \tag{4}$$

where $\boldsymbol{\lambda}_{k,i}$ corresponds to the set of parameters of the approximate posterior distribution $q(\boldsymbol{z}_{k,i}|\boldsymbol{\lambda}_{k,i})$. The expectation operator is defined as $\mathbb{E}_{q_\lambda} f(z) = \int f(z)q(\boldsymbol{z}|\boldsymbol{\lambda})dz$. Note that the evidence (marginal likelihood) $p(D_{k,i})$ does not depend on $\boldsymbol{z}_{k,i}$. Hence, the negative of the expectation term forms a lower bound on the log evidence since the KL distance is always non-negative. This function is known as evidence lower bound (ELBO) and it is the objective function that is maximized in variational EM. The ELBO has the following form:

$$\mathcal{L} = \sum_{i=1}^{I}\sum_{k=1}^{K} \mathbb{E}_{q_{\lambda_{k,i}}}\Big[\log p(\boldsymbol{z}_{k,i},D_{k,i}) - \log q(\boldsymbol{z}_{k,i}|\boldsymbol{\lambda}_{k,i})\Big], \tag{5}$$

where the first term in the expectation corresponds to the joint distribution of the latent variable $\boldsymbol{z}_{k,i}$ and the associated observed data $D_{k,i}$. The second term corresponds to log of the approximate posterior distribution. The joint distribution has the following form:

$$\log p(\boldsymbol{z}_{k,i},D_{k,i}) = \sum_{l=1}^{L} \log p(\boldsymbol{x}_{kl,i}|\boldsymbol{z}_{k,i}) + \log p(\boldsymbol{x}_{kl,i}|\boldsymbol{z}_{k,i}). \tag{6}$$

The expressions for $p(\boldsymbol{x}_{kl,i}|\boldsymbol{z}_{k,i})$ and $p(\boldsymbol{x}_{kl,i}|\boldsymbol{z}_{k,i})$ are given in Eq. 2 and Eq. 1, respectively. We approximate the posterior distribution of $\boldsymbol{z}_{k,i}$ as Gaussian with the following form:

$$q(\boldsymbol{z}_{k,i}|\boldsymbol{\lambda}_{k,i}) = \mathcal{N}(\boldsymbol{z}_{k,i}|\boldsymbol{m}_{k,i},\boldsymbol{S}_{k,i}), \tag{7}$$

where $\boldsymbol{\lambda}_{k,i} = \{\boldsymbol{m}_{k,i},\boldsymbol{S}_{k,i}\}$ is the set of free parameters. Specifically, $\boldsymbol{m}_{k,i}$ is the posterior mean and $\boldsymbol{S}_{k,i}$ is the posterior covariance. The expectation of the approximate posterior distribution in Eq. 5 corresponds to the Gaussian entropy function, which has a closed form expression. However, the expectation of the joint distribution is intractable to compute. Next, we present an approximation to resolve the issue.

2) An upper bound on the LSE: To see why the conditional expectation is intractable, note that the explicit form of the log likelihood of $\boldsymbol{x}_{kl,i}$ is a multinomial distribution:

$$\log p(\boldsymbol{x}_{kl,i}|\boldsymbol{z}_{k,i}) = \sum_{d=1}^{D} \boldsymbol{x}_{kl,id}(\Theta_{kl,d}\boldsymbol{z}_{k,i} - \mathrm{lse}(\Theta_{kl}\boldsymbol{z}_{k,i})). \tag{8}$$

Taking expectation corresponds to integrating out Gaussian distributed $\boldsymbol{z}_{k,i}$. The conditional expectation of the first term is easily determined since it linearly depends on $\boldsymbol{z}_{k,i}$. However, the expectation of the second term, which requires integrating $\boldsymbol{z}_{k,i}$ over the lse function, is intractable to compute in a closed form. To overcome this issue, we perform quadratic surrogate optimization transfer, in which a quadratic approximation to the lse function [8] is applied. This results in an upper bound on the multinomial log likelihood. This approximation uses the second order Taylor series expansion with a fixed Hessian matrix. Particularly, the quadratic upper bound takes the following form:

$$\mathrm{lse}(\Theta_{kl}\boldsymbol{z}_{k,i}) \leq \frac{1}{2}\boldsymbol{z}_{k,i}^T\Theta_{kl}^T\boldsymbol{A}_l\Theta_{kl}\boldsymbol{z}_{k,i} - \boldsymbol{b}_{kl,i}^T\Theta_{kl}\boldsymbol{z}_{k,i} + \boldsymbol{c}_{kl,i}, \tag{9}$$

where

$$\boldsymbol{A}_l = 0.5[\boldsymbol{I}_{D_{xl}} - (1/(D_{xl}+1))\mathbf{1}_{D_{xl}}\mathbf{1}_{D_{xl}}^T] \tag{10}$$

is a constant Hessian matrix, whose entries depend only on the dimension of the observation space. The other intermediate parameters $\boldsymbol{b}_{kl,i}$ and $\boldsymbol{c}_{kl,i}$ are given as follows:

$$\boldsymbol{b}_{kl,i} = \boldsymbol{A}_l\Phi_{kl,i} - \mathcal{S}(\Phi_{kl,i}), \tag{11}$$

$$\boldsymbol{c}_{kl,i} = \frac{1}{2}\Phi_{kl,i}^T\boldsymbol{A}_l\Phi_{kl,i} - \mathcal{S}(\Phi_{kl,i})^T\Phi_{kl,i} + \mathrm{lse}(\Phi_{kl,i}), \tag{12}$$

where $\Phi_{kl,i}$ is the Taylor series expansion point, which is optimized as a free variational parameter. Note that intermediate parameters are deterministic function of $\Phi_{kl,i}$. Plugging the approximation in Eq. 9 to Eq. 5 results in a convex lower bound on ELBO, denoted as $\mathcal{L}'$, which is $\leq \mathcal{L}$ and tight at $\Phi_{kl,i}$. Using $\mathcal{L}'$ resolves the intractable integration in Eq. 5, resulting in closed form posterior parameter estimates, as described in the next section.

3) Posterior Distributions - E-step: The E-step in the variational EM algorithm computes approximate posterior distributions of the latent variables, which are subsequently used to compute the expectations in Eq. 5. Particularly, there are two parameters to be estimated for each latent variable $\boldsymbol{z}_{k,i}$, which are the mean vector $\boldsymbol{m}_{k,i}$ and the covariance matrix $\boldsymbol{S}_{k,i}$. It is straightforward to maximize over these parameters by using the completing-the-square approach [4] (See Appendix A). The terms that quadratically depend on $\boldsymbol{z}_{k,i}$ in the joint log-likelihood yield the posterior covariance update:

$$\boldsymbol{S}_{k,i} = \Big[\Sigma_k^{-1} + \sum_{l=1}^{L} N_{kl,i}\Theta_{kl}^T\boldsymbol{A}_l\Theta_{kl}\Big]^{-1}, \tag{13}$$

where $N_{kl,i}$ is the total number of counts of the ith data sample. Similarly, the terms that linearly depend on $\boldsymbol{z}_{k,i}$ yield the posterior mean update:

$$\boldsymbol{m}_{k,i} = \boldsymbol{S}_{k,i}\Big[\Sigma_k^{-1}\boldsymbol{\mu}_k + \sum_{l=1}^{L}(\boldsymbol{x}_{kl,i} + N_{kl,i}\boldsymbol{b}_{kl,i})\Theta_{kl}\Big]. \quad (14)$$

Lastly, we update the Taylor series expansion point as:

$$\Phi_{kl,i} = \Theta_{kl}\boldsymbol{m}_{k,i}. \quad (15)$$

Note that the update of $\Phi_{kl,i}$ depends on the posterior mean. Hence, the algorithm repeats the updates in Eq. 13, Eq. 14, and Eq. 15, respectively, until convergence of the expansion point $\Phi_{kl,i}$.

4) Point Estimates - M-step: The M-step in the variational EM algorithm maximizes the ELBO with respect to the model parameters. Using the posterior distributions computed in the E-step, we compute the lower bound $\mathcal{L}^{'}$ by taking the expectations with respect to the posterior distributions. Afterwards, taking the derivatives with respect to the model parameters yields closed form update equations for the model parameters. Specifically, the updates for each Θ_{kl} are given as follows:

$$\Theta_{kl} = \Bigg[\sum_{i=1}^{I_k}(\boldsymbol{x}_{kl,i} + N_{kl,i}\boldsymbol{b}_{kl,i})\boldsymbol{A}_l^{-1}\boldsymbol{m}_{k,i}\Bigg] \Bigg[\sum_{i=1}^{I} N_{kl,i}(\boldsymbol{m}_{k,i}\boldsymbol{m}_{k,i}^T + \boldsymbol{S}_{k,i})\Bigg]^{-1}. \quad (16)$$

The update equations for the mean parameter and covariance of the prior distribution of $\boldsymbol{z}_{k,i}$ then follow as:

$$\boldsymbol{\mu}_k = \frac{1}{I_k}\sum_{i=1}^{I}\boldsymbol{m}_{k,i}, \quad (17)$$

$$\Sigma_k = \frac{1}{I}\sum_{i=1}^{I}(\boldsymbol{m}_{k,i} - \boldsymbol{\mu}_k)(\boldsymbol{m}_{k,i} - \boldsymbol{\mu}_k)^T + \boldsymbol{S}_{k,i}, \quad (18)$$

respectively. The variational EM algorithm is summarized in Algorithm 1.

Note that the model parameters of the proposed model are unidentifiable. Due to isotropic Gaussian prior on the latent variables, arbitrary rotation on Θ_{kl} results in same likelihood [33]. This makes direct interpretation of the inferred latent variables ambiguous. Fortunately, this does not affect the predictive performance nor the predictor of covariance, which are the main focus of this paper.

Algorithm 1 Proposed Variational EM algorithm

Input: $\{D_k\}_{k=1:K}$
Initialize $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \{\boldsymbol{\Theta}_{kl}, \{\boldsymbol{\Phi}_{kl,i}\}_{i=1:I_k}\}_{l=1:L}\}_{k=1:K}$
while not $\mathcal{L}^{'}$ converged **do**
  for $k = 1$ to K **do**
    for $i = 1$ **to** I_k **do**
      Infer posterior covariance $\boldsymbol{S}_{k,i}$ by Eq. 13
      Infer posterior mean $\boldsymbol{m}_{k,i}$ by Eq.14
      for $l = 1$ **to** L **do**
        Update variational parameter $\boldsymbol{\Phi}_{kl,i}$ by Eq. 15
      end for
    end for
    for $l = 1$ **to** L **do**
      Estimate Θ_{kl} by Eq. 16
    end for
    Estimate $\boldsymbol{\mu}_k$ by Eq. 17
    Estimate $\boldsymbol{\Sigma}_k$ by Eq. 18
  end for
  Compute $\mathcal{L}^{'}$ by Eq. 5
end while

D. Model-predicted Density

A predictor of the population covariance matrix can be extracted from the inferred model. There are two fundamental choices in the proposed model that pave the way to a predictor of the covariance matrix of the transcriptomes from the abundance data. First, we select Gaussian latent space that is common for all species, which models the observation covariance matrix with a low-rank decomposition due to inherent low-dimensional latent space. Second, we adopt a quadratic lower bound on the multinomial likelihood, hence the marginal likelihood of the observations can be approximated with a Gaussian distribution whose covariance matrix reveals the correlations between genomes. Particularly, we define a transformed version of the sample $\boldsymbol{x}_{kl,i}$ as $\tilde{\boldsymbol{x}}_{kl,i}$ with the following function:

$$\tilde{\boldsymbol{x}}_{kl,i} = \boldsymbol{A}_l^{-1}(\boldsymbol{b}_{kl,i} + \boldsymbol{x}_{kl,i}), \quad (19)$$

where $\boldsymbol{A}_l$ is the matrix defined in Eq. 10. Then, it is straightforward to show that the likelihood of the transformed data $\tilde{\boldsymbol{x}}_{kl,i}$ is given as follow:

$$\begin{aligned} &p(\tilde{\boldsymbol{x}}_{kl,i}|\Theta_{kl}, \boldsymbol{\mu}_k, \Sigma_k) \\ &\quad = \int \mathcal{N}(\tilde{\boldsymbol{x}}_{kl,i}|\Theta_{kl}\boldsymbol{z}_{k,i}, \boldsymbol{A}_l^{-1})\mathcal{N}(\boldsymbol{z}_{k,i}|\mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)d\boldsymbol{z}_{k,i} \\ &\quad = \mathcal{N}(\tilde{\boldsymbol{x}}_{kl,i}|\Theta_{kl}\boldsymbol{\mu}_k, \boldsymbol{A}_l^{-1} + \Theta_{kl}\Sigma_k\Theta_{kl}^T), \end{aligned} \quad (20)$$

where the covariance matrix $\boldsymbol{C}_{kl,\text{intra}} = \boldsymbol{A}_l^{-1} + \Theta_{kl}\Sigma_k\Theta_{kl}^T$ and the mean vector $\phi_{kl} = \Theta_{kl}\boldsymbol{\mu}_{\boldsymbol{k}}$ are of interest to us, in which $\boldsymbol{C}_{kl,\text{intra}}$ captures intra-species correlations of species l in condition k. To obtain inter-species correlations, define $\tilde{\boldsymbol{A}}^{-1} = \text{diag}(\boldsymbol{A}_1^{-1}, \ldots, \boldsymbol{A}_L^{-1})$ and $\tilde{\Theta}_k = [\Theta_{k1}, \ldots, \Omega_{kL}]$, then $\boldsymbol{C}_{k,\text{inter}} = \tilde{\boldsymbol{A}}^{-1} + \tilde{\Theta}_k\Sigma_k\tilde{\Theta}_k^T$ gives a covariance matrix for both inter-species and intra-species. To convert any covariance matrix to a

proper correlation matrix, which is useful for visualization and analysis, one can use the transformation $\text{Corr} = \text{diag}(\boldsymbol{C})^{-1/2}\boldsymbol{C}\text{diag}(\boldsymbol{C})^{-1/2}$.

E. Computational Complexity

The computational complexity of the variational EM algorithm determines the algorithm's scalability to large datasets. For notational simplicity, we assume that there is only one discrete condition, hence we use I instead of I_k. In the E-step, Eq. 13 computes posterior covariance, which requires multiplication of a $d_z \times d_l$ matrix with its transpose resulting $O(d_z^2 d_l)$ complexity. This process is repeated for each species resulting in $O(Ld_z^2 d_l)$. Inverting the matrix for each sample costs $O(Id_z^3)$. Hence, overall asymptotic complexity for the posterior covariance computation is $O(I(d_z^3 + Ld_z^2 d_l))$. The posterior mean computation in Eq. 14 involves matrix-vector multiplications that require $O(Ld_z d_l)$, and $O(d_z^2)$ due to covariance posterior covariance multiplication. Hence, the total cost per sample is $O(Ld_z d_l + d_z^2)$ and the overall cost is $O(I(Ld_z d_l + d_z^2))$. Consequently, the complexity of the E-step is $O(I(Ld_z d_l + d_z^2 + d_z^3 + Ld_z^2 d_l))$. Removing non-dominant terms results in $O(I(d_z^3 + Ld_z^2 d_l))$. One can see that this scales linearly in terms of L, d_l, and I. On the other hand, the dominant computation in the M-step is for Θ_{kl}. Eq. 16 comprises two terms. The first term requires $O(Id_l d_z)$ due to I times vector-vector outer products. The second term requires $O(Id_z^2 + d_z^3)$ due to vector-vector outer products and subsequently matrix inversion. Multiplying these terms costs $O(d_l d_z^2)$, hence resulting total complexity of $O(L(Id_l d_z + Id_z^2 + d_z^3 + d_l d_z^2))$ for all $l = 1 : L$. It is also clear that this computation scales linearly in terms of L, d_l, and I. Modeling the conditions independently also induces linear complexity in terms of K. In summary, both E and M steps scale linearly in terms of K, L, d_l, and I, which suggests that the proposed optimization algorithm is scalable for large datasets as long as the latent space dimension d_z is relatively small.

III. Experiments

In this section, we perform numerical experiments to illustrate the proposed model. We start with simulation studies, then conclude with experiments on a bacterial microbiome dataset.

A. Simulations

We generate synthetic datasets i) to explain the model selection strategy, ii) to demonstrate the accuracy of the latent embeddings, and iii) to show the ability to capture the covariance structure from observed data.

1) Model Selection: The proposed algorithm estimates the covariance matrix with a low-rank decomposition. The rank of the matrix is equal to the number of components d_z in the latent space. which is a model hyper-parameter to be determined. We use the Bayesian Information Criterion (BIC) to estimate this parameter. The BIC arises

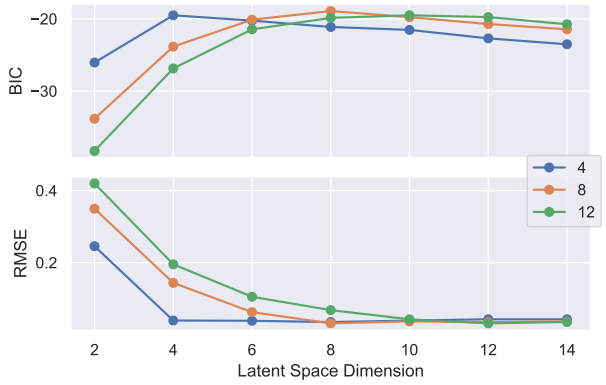


Fig. 2: BIC approximation to the evidence and RMSE of the predicted covariance matrix with respect to the latent space dimension d_z. True dimensions are 4, 8, and 12. Blue, orange, and green curves show RMSE and the BIC penalized log likelihood (BIC), respectively. Note that the BIC exhibits a clear maximum over latent space dimension d_z. BIC values are scaled by factor 10^{-3}.

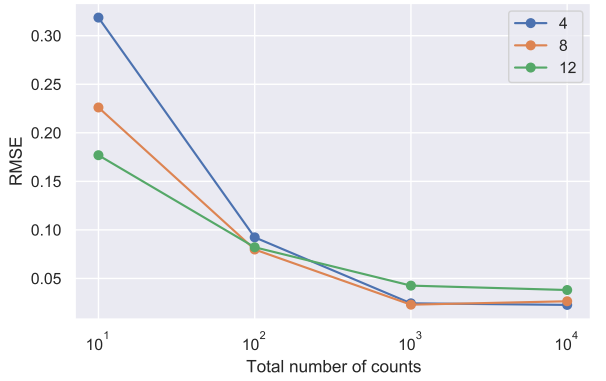


Fig. 3: RSME of the covariance estimation with respect to average total number of counts observed in the metatranscriptomic data. As the counts increase the errors decrease until the counts reach a saturation limit.

from the Laplace approximation to the model posterior $p(M|D_k)$ [21], where M is the complete model including the latent dimension d_z. This results in a Bayesian estimate of d_z: $d_z = \text{argmax}_{d_z}(\log p(D_k) - \text{BIC}/2)$, where BIC is a function of the total number of unknown parameters, which penalizes the log likelihood with a model complexity penalty term. In the proposed model, we use ELBO lower bound to the the likelihood by following [2]. The unknown parameters of the model are $\{\Theta_{kl}\}_{l=1}^{L}$, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma_k}$. Hence, the total number of parameters is $\text{dof} = K \times (d_l + K)$, which is used in the BIC expression as $0.5 \times \text{dof} \times \log(I_k)$. To illustrate the BIC model selection for the proposed model, we simulate three datasets with true latent space dimensions 4, 8, and 12, respectively, and then train multiple models while varying the dimensions d_z over $\{2, 3, \ldots, 12\}$ as the search range. We repeat the experiment 10 times to report the performance. The panel

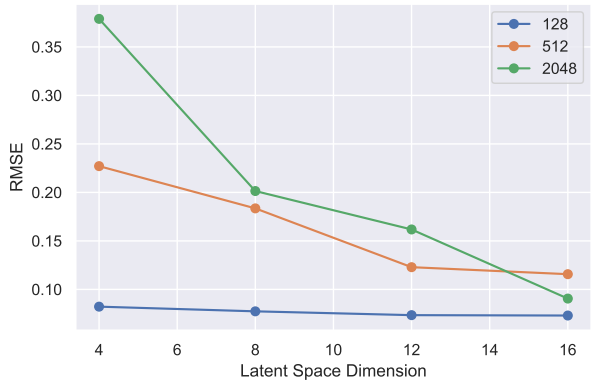
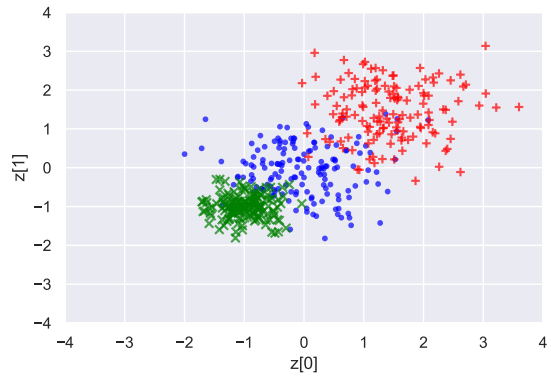
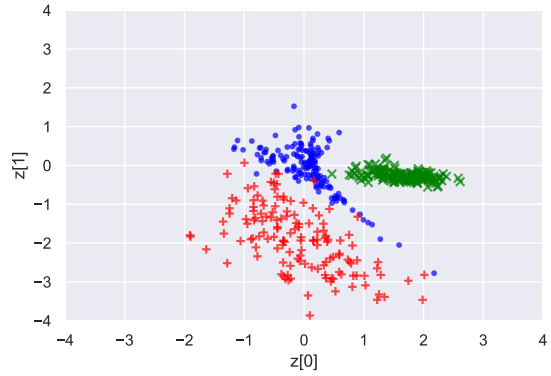


Fig. 4: RMSE of the predicted covariance matrix with respect to the latent space dimension for three different observation space dimensions.



(a) True embeddings



(b) Proposed model embeddings

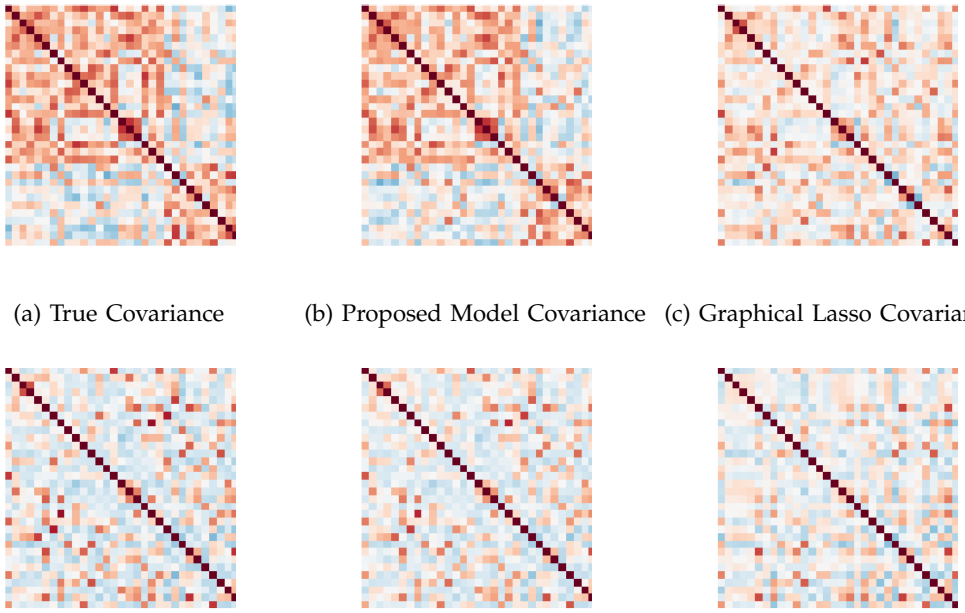
Fig. 5: 2D Latent space visualization of 100D count vectors

on top of Fig. 2 shows the average BIC values obtained after convergence of the variational EM algorithm. We see that maximum BIC is obtained on the vicinity of the true ranks for all the datasets. On the other hand, the panel on the bottom shows the RMSE values of the estimated covariance matrices. One can see that the lowest error is achieved by maximizing d_z, which is likely to result in over-fitting the model. This over-fitting is mitigated by using BIC estimate of d_z.

2) Embedding Characteristics: We generate a synthetic dataset with a 2 dimensional latent space having 3 different classes, i.e., experimental conditions, according to the model specification in Section II-B. The latent variables are sampled for each class from different Gaussian distributions. The associated means are predefined as $[0,0], [1.5,1.5], [-1,-1]$ and the variances of the isotropic covariances are selected as $0.5, 0.5, 0.1$, respectively. Three class conditional densities are generated with different affine transformation parameters. The observation space is 25 dimensional. The observations are sampled from the conditional multinomial distributions with soft-max link function as in Eq. 2. We generate 200 observations for each class with fixed total number of counts, which is 100, per observation, then stack all the observations. Fig. 5.a shows the true embeddings of the resulting dataset. We trained the proposed algorithm with the true latent space dimension. Fig. 5.b shows the embeddings of the model, which are obtained through the posterior distributions. Due to non-identifiability of the model, the latent variables can only be recovered up to a rotation. The distorted shape of the latent clusters in Fig. 5.b is due to the use of the soft-max link function. If there is a large component in the affine transformed latent vector, the other components are washed out, hence such points would map to very close points in the observation space. Notwithstanding the differences between Fig. 5.a and Fig. 5.b, the model preserves the clustering structure accurately.

3) Influence of the Total Counts and Dimensions on Performance: The number of counts of the observed vector $\boldsymbol{x}_{kl,i}$ is an observation-specific parameter, which affects the accuracy of the proposed algorithm. Figure 3 shows the effect of the number of counts $N_{kl,i}$ on the RMSE values of the covariance estimator under three different latent dimension settings. We sample the total counts of a simulated vector from the Poisson distribution with fixed mean. We also fix the observation dimension to 128. RMSE is reported based on averaging 20 experiments. Figure 3 shows that increasing the mean number of counts improves performance. In particular, we see that the total counts $N_{kl,i}$ and the mean error are inversely proportional. This is expected since the number of counts directly affects the posterior uncertainty (Eq. 13) and mean (Eq. 14). The contribution to the ELBO of the observations increases as the total count increases. Furthermore, for low number of counts, the covariance matrix becomes harder to predict due to higher vulnerability to over-fitting. On the other hand, Fig. 4 demonstrates the opposite trend when the dimension of the observation space dimension is increased. Here the total mean counts is fixed to 100. In higher dimensional datasets, the model struggles to estimate the covariance structure when the rank is low. However, this phenomenon diminishes when we observe more counts as can be seen in Fig. 3.

4) Baseline Algorithms: Here we show performance comparisons of the proposed method relative to four baseline



(a) True Covariance (b) Proposed Model Covariance (c) Graphical Lasso Covariance

(d) Emprical Covariance (e) Ledoit-Wolf Covariance (f) FA Covariance

Fig. 6: Estimated covariance matrices produced by the considered algorithms for a three species community with simulated transcript data. For more details on the model see Section III-A5. The proposed model provides a much more accurate estimated covariance than than do the other methods.

methods for estimating the underlying covariance and inverse covariance matrices. i) Empirical covariance, which is the sample covariance. ii) The Ledoit-Wolf estimator [24], which uses shrinkage regularization to perform MAP estimation for the covariance matrix by assigning an inverse Wishart prior on the covariance matrix. iii) Gaussian Copula GraphicalLasso [26], which penalizes the covariance matrix in a different way, particularly, with L1-norm constraints on the precision matrix after transforming the data by using Gaussian copulas. Regularization forces the entries of the precision matrix to be sparse. iv) Factor Analysis [33] uses another form of regularization of the covariance matrix by imposing low-rank structure. Each of these baseline methods is expected to perform best when the data, or its transformed version, is normally distributed. For the Ledoit-Wolf, GraphicalLasso, and FA, as is customary, we first normalize the data by subtracting the mean and dividing by the variance, before running these methods. On the other hand, as the non-Gaussian counting nature of the data is explicitly modeled in our proposed model, our algorithm is run on the raw observations. For model selection in both FA and proposed model, we use the exact rank of the simulated dataset. The regularization coefficient of Gaussian Copula GraphicalLasso algorithm is estimated by using 5-fold cross-validation. For the Ledoit-Wolf algorithm, we used the expression for the shrinkage coefficient given in [24]. We use `scikit-learn` implementations of the baseline methods.

5) Simulating Model Communities: Next, we generate a synthetic dataset which contains the transcript abundance data (an estimate of gene expression) of two different species existing in the same community, hence $L = 2$. The latent variables $\boldsymbol{z}_i$ with dimension $d_z = 5$ are generated for each measurement site by sampling from $\boldsymbol{z}_i \sim \mathcal{N}(\mathbf{0}_{d_z}, \boldsymbol{I}_{d_z})$, where i indexes the replicate for $i = 1 : I$. These latent variables have elements that correspond to the hidden factors generating the data, such as environmental variables, mediator species effects, and direct associations. We transform the latent variables to the probabilities in the observation space, whose dimensions (abundance of transcripts) are chosen as $d_1 = 20, d_2 = 10$ by using affine and subsequently soft-max transformations as described in Section II-B. The parameters $\Theta_l \in \mathbb{R}^{d_l \times d_z}$, are chosen randomly by sampling from a zero mean multi-variate normal distribution. Then, we sample the observed data $\boldsymbol{x}_{l,i}$ from the multinomial distribution. The total counts $N_{l,i}$ of a sample is chosen randomly by sampling from a Poisson distribution with rate parameter 1000. We simulate a total of $I = 200$ replicates for each dataset. The true covariance matrix is then given as $\tilde{\Theta}_l \tilde{\Theta}_l^T$, where $\tilde{\Theta}_l = [\Theta_1, \Theta_2]$.

6) Correlation Results: Fig. 6 show the estimated covariance matrices of the baseline algorithms, the proposed algorithm, alongside with the ground truth matrix, when simulated dataset is realized. The proposed model can recover the covariance structure accurately. The relatively poorer accuracy of the other methods can be attributed

	Algorithm	Covariance	Precision
Cov	Empirical	.231 ± .013	.510 ± .021
	Ledoit-Wolf	.225 ± .011	.150 ± .004
	FA	.237 ± .014	.086 ± .003
	GLasso	.173 ± .013	.073 ± .002
	Proposed	.096 ± .022	.003 ± .001

TABLE I: Mean and standard deviation of RMSE between the estimated covariance matrices and ground truth over 10 different realizations of the simulated abundance dataset.

to several factors. First, these models do not exploit the counting nature of the data. The second reason is that the covariance matrix is simulated with low-rank structure, which is not taken into account by the Gaussian Copula Graphical-Lasso, Ledoit-Wolf, or standard sample covariance estimation methods. As the data was simulated from the proposed model, the proposed algorithm naturally performs better. Note also that, for multiple species, the proposed model can discover both inter-species and intra-species correlations. Table I shows the resulting RMSE values between the estimated and the ground truth covariance matrices for the aforementioned simulation setting. The proposed model achieves lower error in overall. This is expected since the model uses an ELBO approximation to the true marginal likelihood function.

B. Bacterial Community Experiment

In this section, we demonstrate a real world use-case of the proposed model: transcript analysis of a bacterial model community called THOR [17].

Microbial model communities are useful to understand principles that govern community behaviours [3], [14], [12], [42]. The Hitchhikers Of the Rhizosphere (THOR) is a model community consisting of three microbial species, *Bacillus cereus, Flavobacterium johnsoniae*, and *P. koreensis* that co-isolate from field-grown soybean roots. The organisms in THOR represent three dominant rhizosphere taxa (at the phylum level), and are common in soil and the mammalian gut. *B. cereus* is a Firmicute that carries *F. johnsoniae*, a member of the Bacteriodetes, and *P. koreensis*, a member of the Proteobacteria, as hitchhikers [27]. Due to their abundance in several environments, their may demonstrated interactions in the lab and field, and their genetic tractability, these species make a useful model community with relevance to the natural world. The model community provides a simple system in which to study and model community level interactions, which are poorly understood. Developing governing principles of community behavior may lead to strategies to manipulate microbiomes for human or environmental health.

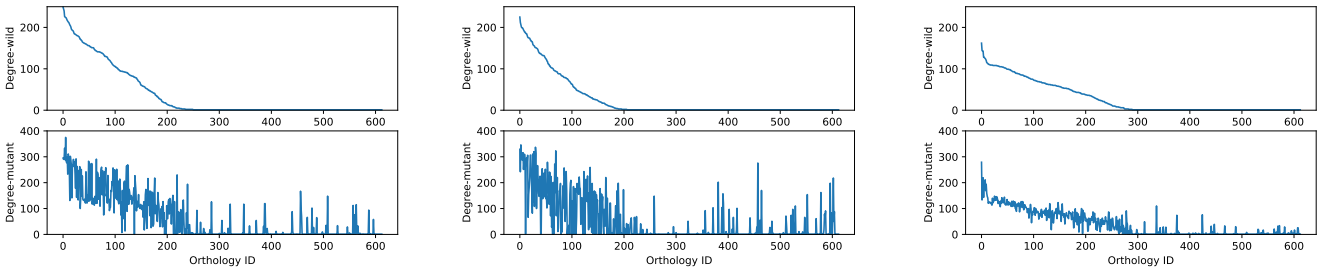
The dataset is collected under two conditions associated with the treatments applied to *P. koreensis*. In the first condition the THOR community contains the wild type *P. koreensis* strain and in the second condition the wildtype is replaced with a mutant of *P. koreensis* that does not produce koreenceine antibiotics. Production of koreenceines is an important factor in community interactions because they inhibit growth of *F. johnsoniae* [28] and *B. cereus* protects *F. johnsoniae* by modulating koreenceine levels. By using our proposed model, in particular the associated estimated joint probability density of the data, we will be able to reveal effects of the treatment. Since the joint probability density model is parameterized by the mean and covariance of a multivariate Gaussian latent variable (See Section II-D), the mean and covariance parameters play the principal role in our metatranscriptomic analysis. For brevity we focus our discussion on the inferred covariance parameters here (See Supplementary for discussion of the mean parameters inferred by the model).

The microbial community dataset consists of a total of 17244 gene transcripts associated with three species. There were respectively 38 and 36 replicates for the community with wildtype and mutant strains of *P. koreensis*. 343 transcripts were removed from the analysis as they had zero counts over all experimental replicates. After removing these transcripts, *B. cereus, F. johnsoniae,* and *P. koreensis* express 5903, 5146, 5852 transcripts, respectively. We reduced the dimension of the feature space using orthological groupings of gene transcripts into metabolic pathways[3]. Specifically, after pathway mapping each feature corresponds to a transcritpional orthology ID, and the associated data is the summation of the counts of the transcripts tagged with that ID. We aggregated all the transcripts that were not mapped to any Kegg ortholog into a single non-assigned orthology ID, denoted KXXXXX, and we only considered those ortholog IDs that are present in all 3 species. This filtering resulted in a set of 613 ortholog IDs, which corresponds to the dimension of the feature space used in our model.

The rank of the proposed model was determined by successively fitting the model to latent spaces of dimensions ranging between 5 and 50 with increments of 5. Then, the optimal model rank was determined as the latent dimension that yields the highest value of the BIC as described in Section III-A1. The optimal model rank was found to be 40. The parameters (mean and covariance) of the models were subsequently refitted with the optimal dimension. The probability distribution of the data is computed under the wildtype and mutant conditions, whose explicit form is given in Eq. 20 as a marginalization over the latent variables.

Network centrality changes: We evaluate the effect of removal of koreenciene (mutant) on the centrality of the inferred 613×613 correlation network of metabolic pathways. Here the centrality of a vertex of the network is measured by vertex degree, i.e., the number of edges connecting the vertex. To ensure that the networks contain only the most biologically significant edges in the networks, we applied a very high correlation threshold (0.95) to the respective inferred wild-type and mutant

[3]The transcriptional orthology mappings of the THOR gene transcripts to metabolic pathways were obtained using Kegg https://www.genome.jp/kegg/. See supplementary for an example.



(a) B. cereus (b) F. johnsoniae (c) P. koreensis

Fig. 7: Effect of koreenciene removal on the the centrality (vertex degree) of vertices in the transcriptional orthology correlation networks inferred from our model for the experimental THOR dataset. For each species, the ortholog IDs are sorted in decreasing order of the wildtype vertex degree. The upper row shows plots of the degree of each vertex (transcriptional orthology ID), in descending order of magnitude, for the wildtype condition. The bottom row shows corresponding plots of the vertex degree when the koreenceine pathway is removed (mutant condition), under the same ordering of vertices as in the top row. *P. koreensis* preserves its network connectivity better than the other two species. The network connectivity of *F. johnsoniae* is the most affected by koreenciene removal.

correlation matrices produced by fitting our proposed graphical model to the data. Using such a high threshold is in line with established RNA-Seq network inference practices [25]. Figure 7 illustrates the effect of removal of koreenceine on the degrees of the nodes (transcriptional orthology IDs) in these networks. Comparison of the upper panels with the lower panels of the figure indicates that the vertex degree distribution of *F. johnsoniae* is most affected, followed by *B. cereus*, with *P. koreensis* the least affected. This relative ordering of sensitivity of the three species to koreenceine removal shown for vertex degree in Fig. 7 mirrors the relative ordering of sensitivity shown for the mean changes (See Fig. 2 and associated discussion in the Supplementary).

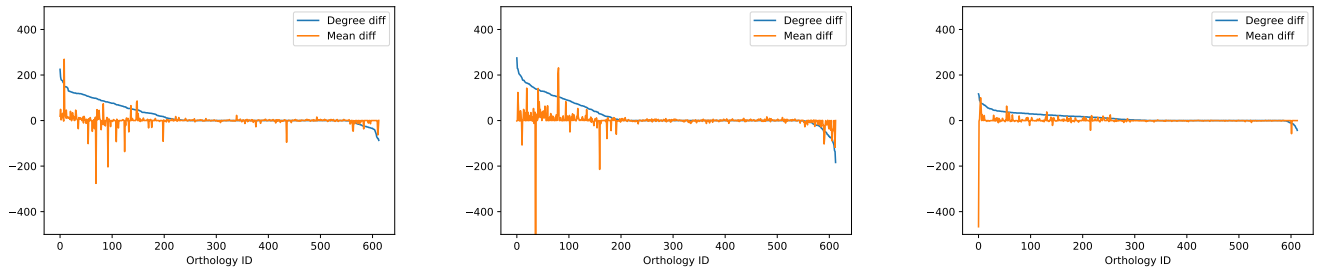
Fig. 7 illustrates the relative effect of koreenceine removal on increases vs decreases in vertex degree of the transcriptional orthology correlation network for each species. In the figure the transcriptional orthology IDs are sorted according to the difference between mutant vs wildtype vertex degree. The blue curve shows the resultant vertex degree difference and the orange curve shows the vertex mean difference. Observe that the order of decreasing differences of vertex degree does not correspond to the order of decreasing differences in vertex mean. However, a change in the vertex mean almost always accompanies a change in vertex degree, although the converse is not true. Also note from the asymmetry of the blue curves in Fig. 7 that the mutant's networks have many more vertices that increase than decrease in vertex degree as compared to the wildtype. Thus koreenceine removal seems to increase network centrality of a large number of transcriptional orthologs, especially for *F. johnsoniae*. We point out that the large spikes that appear in the orange curves (vertex mean difference) for *F. johnsoniae* and *P. koreensis*, correspond to the ID KXXXXX, which are genes that were not mapped to any Kegg transcriptional ortholog. Further discussion can be found in the supplementary.

In summary, the proposed model can provide two important data analysis components for microbiome model community analysis. First, we can assess transcriptional orthology composition changes under the treatment by observing the means of the marginal distributions provided by the proposed model. Second, we can assess the second order interaction changes by using the correlation networks that are obtained from the covariance matrices of the marginal distributions. These two components along with the abundance ratio analysis in [17] provide a complementary analysis of microbial model communities, which can further be interpreted by microbiologists.

IV. Conclusion

A hierarchical Bayesian latent variable model was proposed for the joint analysis of multiple discrete datasets. We explained the associations between the features of the datasets with a common lower dimensional latent space, represented by a set of independent identically distributed Gaussian random variables. To overcome the lack of conjugacy between the multinomial observation distribution and the Gaussiam latent space distribution, we developed a variational EM algorithm based on quadratic bound approximations for estimating the parameters in the model. The computation of the algorithm scales linearly with the number of features, samples, and datasets. Simulation studies show that the proposed model can recover low-rank covariance structures accurately. Furthermore, our real-world microbiome experiment demonstrates the potential real-world utility of the model for exploration of correlation and associated networks for dichotomous microbiome data.

There are several promising directions for future work. One possible area of future work is to incorporate system dynamics into the latent space so as to explicitly capture temporal correlations. In particular, there is increasing



(a) B. cereus (b) F. johnsoniae (c) P. koreensis

Fig. 8: Effect of koreenceine removal on vertex centrality and vertex mean counts for the transcriptional orthology correlation network. For each species, the ortholog IDs are sorted in decreasing order of the vertex degree difference between mutant and wildtype. It is notable that, with few exceptions, all orthology IDs with significant changes in vertex mean also have changes vertex degree, but not conversely. Furthermore, the asymmetry of the blue curve suggests that the removal of koreenceine is associated with an increase in network connectivity (many more vertices whose degrees increase than decrease), especially in *F. johnsoniae*.

interest in collecting longitudinal microbiome data for studying adaptation, resilience, and dynamics over time. Incorporation of a state-space dynamical model into our framework can reveal temporal evolution of the interactions between the genomes. Another future direction is to improve the parsimony of the model by incorporating sparsity into the latent representation by using sparsity inducing priors for the covariance or inverse covariance (precision) matrices.

References

[1] Vanessa Aguiar-Pulido, Wenrui Huang, Victoria Suarez-Ulloa, Trevor Cickovski, Kalai Mathee, and Giri Narasimhan. Metagenomics, metatranscriptomics, and metabolomics approaches for microbiome analysis: supplementary issue: bioinformatics methods and applications for big metagenomics data. *Evolutionary Bioinformatics*, 12:EBO–S36436, 2016.

[2] Matthew J Beal and Zoubin Ghahramani. Variational bayesian learning of directed graphical models with hidden variables. *Bayesian Analysis*, 1(4):793–831, 2006.

[3] Johan Bengtsson-Palme. Microbial model communities: To understand complexity, harness the power of simplicity. *Computational and Structural Biotechnology Journal*, 18:3987–4001, 2020.

[4] C.M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006.

[5] David Blei and John Lafferty. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006.

[6] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[7] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[8] Dankmar Böhning. Multinomial logistic regression algorithm. *Annals of the institute of Statistical Mathematics*, 44(1):197–200, 1992.

[9] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[10] Michael Braun and Jon McAuliffe. Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, 105(489):324–335, 2010.

[11] Wray Buntine and Aleks Jakulin. Discrete component analysis. In *International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection"*, pages 1–33. Springer, 2005.

[12] Marc G. Chevrette, Jennifer R. Bratburd, Cameron R. Currie, Reed M. Stubbendieck, and Barbara Methe. Experimental microbiomes: Models not to scale. *mSystems*, 4(4):e00175–19, 2019.

[13] Michael Collins, Sanjoy Dasgupta, and Robert E Schapire. A generalization of principal components analysis to the exponential family. In *Nips*, volume 13, page 23, 2001.

[14] Karen De Roy, Massimo Marzorati, Pieter Van den Abbeele, Tom Van de Wiele, and Nico Boon. Synthetic microbial ecosystems: an exciting tool to understand and apply microbial communities. *Environmental Microbiology*, 16(6):1472–1481, 2014.

[15] Connor R Fitzpatrick, Isai Salas-González, Jonathan M Conway, Omri M Finkel, Sarah Gilbert, Dor Russ, Paulo José Pereira Lima Teixeira, and Jeffery L Dangl. The plant microbiome: From ecology to reductionism and beyond. *Annual Review of Microbiology*, 74:81–100, 2020.

[16] David J Harris. Inferring species interactions from co-occurrence data with markov networks. *Ecology*, 97(12):3308–3314, 2016.

[17] Amanda Hurley, Marc G. Chevrette, Natalia Rosario-Meléndez, Jo Handelsman, and Gerard D. Wright. Thor's hammer: the antibiotic koreenceine drives gene expression in a model microbial community. *mBio*, 0(0):e02486–21, 2022.

[18] Tommi S Jaakkola and Michael I Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.

[19] Mohammad Khan, Shakir Mohamed, Benjamin Marlin, and Kevin Murphy. A stick-breaking likelihood for categorical data analysis with latent gaussian models. In *Artificial Intelligence and Statistics*, pages 610–618. PMLR, 2012.

[20] Mohammad Emtiyaz E Khan, Guillaume Bouchard, Kevin P Murphy, and Benjamin M Marlin. Variational bounds for mixed-data factor analysis. *Advances in Neural Information Processing Systems*, 23:1108–1116, 2010.

[21] Sadanori Konishi and Genshiro Kitagawa. Information criteria and statistical modeling. 2008.

[22] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474, 2017.

[23] Thomas K Landauer and Susan T Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

[24] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

[25] Franziska Liesecke, Dimitri Daudu, Rodolphe Dugé de Bernonville, Sébastien Besseau, Marc Clastre, Vincent Courdavault, Johan-Owen De Craene, Joel Crèche, Nathalie Giglioli-Guivarc'h, Gaëlle Glévarec, et al. Ranking genome-wide correlation measurements improves microarray and rna-seq based global and targeted co-expression networks. *Scientific reports*, 8(1):1–16, 2018.

[26] Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(10), 2009.

[27] Gabriel L. Lozano, Juan I. Bravo, Manuel F. Garavito Diago, Hyun Bong Park, Amanda Hurley, S. Brook Peterson, Eric V. Stabb, Jason M. Crawford, Nichole A. Broderick, Jo Handelsman, Gary M. Dunny, Roberto Kolter, and Irene Newton. Introducing thor, a model microbiome for genetic dissection of community behavior. *mBio*, 10(2):e02846–18, 2019.

[28] Gabriel L. Lozano, Hyun Bong Park, Juan I. Bravo, Eric A. Armstrong, John M. Denu, Eric V. Stabb, Nichole A. Broderick, Jason M. Crawford, Jo Handelsman, and Marie A. Elliot. Bacterial analogs of plant tetrahydropyridine alkaloids mediate microbial interactions in a rhizosphere model system. *Applied and Environmental Microbiology*, 85(10):e03058–18, 2019.

[29] Bob Mau and Michael A Newton. Phylogenetic inference for binary data on dendograms using markov chain monte carlo. *Journal of Computational and Graphical Statistics*, 6(1):122–131, 1997.

[30] Rahul Mazumder and Trevor Hastie. The graphical lasso: New insights and alternatives. *Electronic journal of statistics*, 6:2125, 2012.

[31] Gary W Miller. *The exposome: A primer*. Elsevier, 2013.

[32] Shakir Mohamed, Zoubin Ghahramani, and Katherine A Heller. Bayesian exponential family pca. *Advances in neural information processing systems*, 21:1089–1096, 2008.

[33] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[34] Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023.

[35] Gordana C Popovic, Francis KC Hui, and David I Warton. A general algorithm for covariance modeling of discrete data. *Journal of Multivariate Analysis*, 165:86–100, 2018.

[36] Gordana C Popovic, David I Warton, Fiona J Thomson, Francis KC Hui, and Angela T Moles. Untangling direct species associations from indirect mediator species effects with graphical models. *Methods in Ecology and Evolution*, 10(9):1571–1583, 2019.

[37] F James Rohlf and Robert R Sokal. Comparing numerical taxonomic studies. *Systematic Biology*, 30(4):459–490, 1981.

[38] Migun Shakya, Chien-Chi Lo, and Patrick SG Chain. Advances and challenges in metatranscriptomic analysis. *Frontiers in genetics*, 10:904, 2019.

[39] Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*, 2017.

[40] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[41] Max Welling, Chaitanya Chemudugunta, and Nathan Sutter. Deterministic latent variable models and their pitfalls. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 196–207. SIAM, 2008.

[42] Benjamin E. Wolfe. Using cultivated microbial communities to dissect microbiome assembly: Challenges, limitations, and the path ahead. *mSystems*, 3(2):e00161–17, 2018.

[43] Grace Yoon, Irina Gaynanova, and Christian L Müller. Microbial networks in spring-semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data. *Frontiers in genetics*, 10:516, 2019.

Appendix A
Estimation of Posterior Parameters

Log-likelihood of Multivariate Normal distribution $\log \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be written as:

$$-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + const$$

in which the second order term in $\boldsymbol{x}$ corresponds to the inverse of covariance matrix $\boldsymbol{\Sigma}$, and the linear term corresponds to the mean when multiplied with $\boldsymbol{\Sigma}$. Inferring the mean and covariance from linear and quadratic terms is called completing the square approach. We make use of this method to infer the posterior distributions of $\boldsymbol{z}_{k,i}$, which is denoted as $q(\boldsymbol{z}_{k,i}|\boldsymbol{m}_{k,i}, \boldsymbol{S}_{k,i})$. Given the joint likelihood in Eq. 6 and quadratic approximation in Eq. 9, one can collect the quadratic terms in $\boldsymbol{z}_{k,i}$ as follows:

$$-\frac{1}{2}\boldsymbol{z}_{k,i}^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{z}_{k,i} - \sum_{l=1}^{L}\frac{N_{kl,i}}{2}\boldsymbol{z}_{k,i}^T\Theta_{kl}^T\boldsymbol{A}_l\Theta_{kl}\boldsymbol{z}_{k,i},$$

which follows Eq. 13 for posterior covariance $\boldsymbol{S}_{k,i}$ estimate. Similarly, the linear terms are collected as:

$$\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k\boldsymbol{z}_{k,i} + \sum_{l=1}^{L}\boldsymbol{x}_{kl,i}\boldsymbol{\Theta}_{kl}\boldsymbol{z}_{k,i} + N_{kl,i}\boldsymbol{b}_{kl,i}^T\boldsymbol{\Theta}_{kl}\boldsymbol{z}_{k,i}.$$

Collecting the terms and multiplying with the posterior covariance estimate yields posterior mean $\boldsymbol{m}_{k,i}$ estimate as given in Eq. 14.

B Supplementary

Metabolism

Global and overview maps
01100 Metabolic pathways (1093)

K00001	Bc_ctg1_397, Bc_ctg1_4917, Bc_ctg1_5134, Fj_ctg1_4098
K00003	Bc_ctg1_2360, Bc_ctg1_4276, Pk_ctg1_4531
K00005	Bc_ctg3_122
K00008	Fj_ctg1_4232
K00012	Bc_ctg1_2147, Bc_ctg1_2218, Bc_ctg2_104, Fj_ctg1_368, Pk_ctg1_2576, Pk_ctg1_3377
K00013	Bc_ctg1_3778, Fj_ctg1_2949, Pk_ctg1_4701
K00014	Bc_ctg1_1318, Fj_ctg1_2653, Pk_ctg1_511, Pk_ctg1_5539
K00015	Bc_ctg1_3787, Fj_ctg1_1956
K00016	Bc_ctg1_1853, Bc_ctg1_1965, Bc_ctg1_4233, Bc_ctg3_168
K00018	Pk_ctg1_708
K00019	Bc_ctg1_1014
K00020	Pk_ctg1_4293, Pk_ctg1_4875
K00023	Bc_ctg1_3688
K00024	Bc_ctg1_1569, Fj_ctg1_2314
K00029	Fj_ctg1_1692, Pk_ctg1_5177
K00031	Bc_ctg1_1570, Pk_ctg1_1911, Pk_ctg1_1912
K00032	Pk_ctg1_2877
K00033	Bc_ctg1_279, Bc_ctg1_4564, Fj_ctg1_4935, Pk_ctg1_2521

(a) Functional orthologies of THOR genomes

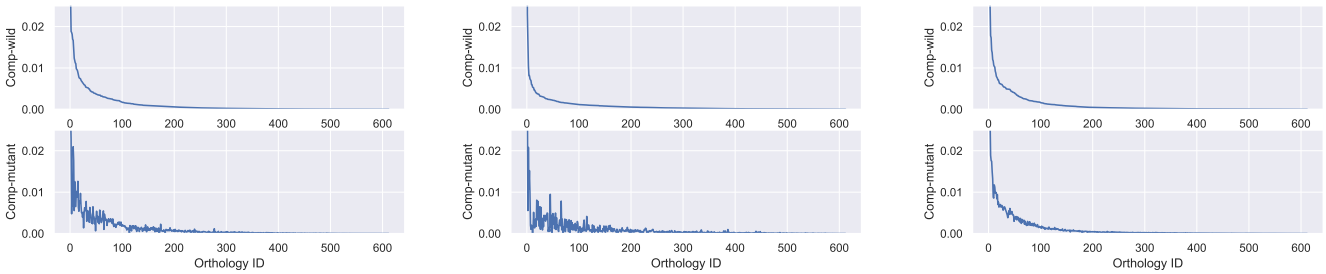
Entry	K00001 KO
Symbol	E1.1.1.1, adh
Name	alcohol dehydrogenase [EC:1.1.1.1]
Pathway	map00010 Glycolysis / Gluconeogenesis
	map00071 Fatty acid degradation
	map00350 Tyrosine metabolism
	map00620 Pyruvate metabolism
	map00625 Chloroalkane and chloroalkene degradation
	map00626 Naphthalene degradation
	map00830 Retinol metabolism
	map00980 Metabolism of xenobiotics by cytochrome P450
	map00982 Drug metabolism - cytochrome P450
	map01100 Metabolic pathways
	map01110 Biosynthesis of secondary metabolites
	map01120 Microbial metabolism in diverse environments
	map01220 Degradation of aromatic compounds

(b) Kegg Orthology Kooo01, an alcohol dehydrogenase

Fig. 9: Examples of classification of THOR transcripts based on metabolic functions inferred from gene sequences. Multiple species may have the same transcriptional orthology. For instance, the Kegg functional orthology Kooo01 is a group of transcipts that encode alcohol dehydrogenase (AD), an enzyme that plays a role in several cellular processes. AD homologues are found in genomes of both *Bacillus cereus* and *Flavobacterium johnsoniae*.

Mean changes over mutant/wildtype conditions. For each species, and under either of the wildtype or mutant conditions, we estimate the 613-dimensional mean vector of Kegg functional orthology counts, as determined by the iterative estimate Eq. (17). We then transform this estimated mean vector to a probability distribution over IDs using the multivariate softmax transformation, yielding what we call the transcriptional orthology probability distribution that quantifies the composition of orthologs for each species under each condition. The change to the species-specific probability distribution caused by the removal of koreenceine is quantified by comparing the wildtype and mutant distributions, shown in Fig. 10. To quantify the effect of koreenceine removal, we evaluate the Hellinger distances between the respective wildtype and mutant transcriptional orthology probability distributions. The distances where found to be 0.12, 0.18, and 0.05 for *B. cereus, F. johnsoniae*, and *P. koreensis,* respectively. This suggests that the transcriptional orthology composition of *F. johnsoniae* is most affected under the mutant condition. In contrast the transcritpional composition of *P. koreensis* changes the least of the three species.

We also observed that the removal of koreenceine had a substantial effect on the relative abundance of different microbial species. For wildtype and mutant conditions the relative abundances of different species in the model community were computed by taking the total number of RNA-Seq counts for each of the 3 species and normalizing it by its sum. The change in relative abundances was quantified by relative change abundance ratio, which are found to be -0.01 (-1%) for *B. cereus*, 0.10 (10%) for *F. johnsoniae*, and -0.08(-8%) for *P. koreensis*, respectively. These quantities are compatible with [17], which suggests that deleting the koreenceine biosynthetic pathway in *P. koreensis* enhances growth of *F. johnsoniae* and does not affect *B. cereus*. It is interesting that while the removal of koreenceine does not affect the abundance of *B. cereus* it has substantial effect on the transcriptional orthology probability distribution, as predicted by our model, discussed in the previous paragraph.



(a) B. cereus (b) F. johnsoniae (c) P. koreensis

Fig. 10: Comparison of the effect of koreenceine removal on the mean transcriptional orthology compositions for each of the three species. The ortholog IDs are sorted in decreasing order of the wildtype's Hellinger distance to the mutant composition distribution. Top panel corresponds to wild-type compositions whereas bottom panel shows the mutant compositions, in which the koreenceine pathway of *P. koreensis* has been removed, with aligned indexes. Hellinger distances between the estimated probability vectors of the treatments are 0.12, 0.18, 0.05 for *B. cereus*, *F. johnsoniae*, and *P. koreensis*, respectively, which suggests that the transcriptional orthology composition of *F. johnsoniae* is most affected by koreenceine removal.

Orthology ID	DegDiff	DegW	DegM	MeanDiff	MeanW	MeanM
Emerging Connections						
K15777 (4,5-DOPA dioxygenase extradiol)	**275**	1	276	-0.08	0.10	0.01
K07646 (OmpR family, sensor histidine kinase KdpD)	**229**	94	323	0.00	0.00	0.00
K03040 (DNA-directed RNA polymerase subunit alpha)	**227**	32	259	6.16	1.68	7.84
Extinguished Connections						
K01610 (phosphoenolpyruvate carboxykinase (ATP))	**-132**	135	3	-3.29	4.64	1.35
K01358 (ATP-dependent Clp protease, protease subunit)	**-136**	137	1	-5.88	6.89	1.01
K00274 (monoamine oxidase)	**-184**	186	2	-0.00	0.00	0.00
Up-regulated Functions						
K02358 (elongation factor Tu)	104	225	329	**11.57**	23.02	34.59
K02355 (Not found in KEGG)	104	189	293	**10.22**	10.59	20.81
K03076 (preprotein translocase subunit SecY)	165	78	243	**7.12**	2.35	9.48
Down-regulated Functions						
KXXXXX (Unannotated)	139	207	346	**-57.08**	536.04	478.95
K04043 (molecular chaperone DnaK)	30	212	242	**-10.73**	16.28	5.55
K01358 (ATP-dependent Clp protease, protease subunit)	-136	137	1	**-5.88**	6.89	1.01
Emerging Connections w/o regulations						
K07646 (OmpR family, sensor histidine kinase KdpD)	**229**	94	323	**0.00**	0.00	0.00
K02238 (Not found in KEGG)	**161**	151	312	**0.00**	0.00	0.00
K00788 (thiamine-phosphate pyrophosphorylase)	**129**	177	306	**0.00**	0.00	0.00
Extinguished Connections w/o regulations						
K00274 (monoamine oxidase)	**-184**	186	2	**-0.00**	0.00	0.00
K03781 (catalase)	**-109**	175	66	**0.00**	0.00	0.00
K02575 (MFS transporter, nitrate/nitrite transporter)	**-50**	150	100	**0.00**	0.00	0.01
Up-regulated Functions w/o degree changes						
K02495 (oxygen-independent coproporphyrinogen III oxidase)	**0**	1	1	**0.81**	0.16	0.98
K03530 (Not found in KEGG)	**0**	1	1	**0.67**	0.22	0.89
K02899 (large subunit ribosomal protein) L27	**1**	1	2	**0.64**	0.91	1.56
Down-regulated Functions w/o degree changes						
K03695 (ATP-dep Clp protease ATP-binding subunit ClpB)	**-2**	3	1	**-1.37**	1.86	0.48
K00265 (glutamate synthase (NADPH) large chain)	**0**	1	1	**-0.73**	0.96	0.22
K03733 (Not found in KEGG)	**0**	1	1	**-0.65**	0.91	0.25

TABLE II: Top 3 transcripts of *Flavobacterium johnsoniae* that change the most according to the eight different treatment effect definitions (emerging connection, extinguished connections, up-regulated functions, down-regulated functions, emerging connections w/o regulations, extinguished connections w/o regulations) . Note that the numbers in the 3 last columns are scaled by 10^3 to reduce clutter in rendering the table. This table provides a snapshot of the transcriptional orthologs in *Flavobacterium johnsoniae* that are the most sensitive to removal of koreenceine under 8 different sensitivity criteria. We highlight *Flavobacterium johnsoniae* here because it is the species that is most affected although analogous CSV tables for other species can be found in the supplementary files. Emerging and extinguished connections correspond to, respectively, positive and negative centrality changes, as measured by vertex degree, in the transcriptional ortholog correlation graph. Up-regulated and down regulated functions correspond to the orthologies whose compositions change positively and negatively, respectively. Emerging and extinguished connections w/o mean change denote transcriptional orthologs for which there are changes in vertex degree but no change in vertex mean. Up-regulated and down regulated w/o degree changes denote transcriptional orthologs that change have changes in mean but not in vertex degree. The proposed model can provide two important data analysis components for microbiome model community analysis. First, we can assess transcriptional orthology composition changes under the treatment by observing the means of the marginal distributions provided by the proposed model. Second, we can assess the second order interaction changes by using the correlation networks that are obtained from the covariance matrices of the marginal distributions. These two components provide a complementary analysis of microbial model communities, which can further be interpreted by microbiologists.