A Graphical Model for Fusing Diverse Microbiome Data

Mehmet Aktukmak®, Haonan Zhu®, Marc G. Chevrete, Julia Nepper®, Shruthi Magesh®, Jo Handelsman®, and Alfred Hero®

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 corresponds 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.

Index Terms—Bayesian probabilistic graphical model, data fusion, microbial data analysis, variational optimization.

I. INTRODUCTION

We introduce a Bayesian graphical model for joint 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: the fusion of microbiome data from diverse microbial communities.

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 [22]. Another area of study is the spectrum of responses of microbiomes to stressors, collectively called the microbial exposome [49].

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 environments. 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 [2], [45], [60]. The resulting datasets consists of species abundance (count of RNA occurrences) for different samples obtained from various communities. This paper introduces a Bayesian graphical model for the metatranscriptomics problem, and 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 interspecies genome-to-genome mappings or knowledge of interspecies transcriptomic pathway correspondences.

The main feature of our model is that it estimates the global covariance structure of gene expression when
observations are in the form of count vectors produced by RNA-Seq. Correlations between transcript abundances are informative about the effect of environmental conditions on microbial communities [63]. 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 a 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], [12] 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) [53] is a classical method that is a generalization of Principal Component Analysis (PCA) [7] and Probabilistic PCA [68]. 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 [35], Multinomial PCA [14], and Latent Dirichlet Allocation (LDA) [10]. 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 [8], [52].

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) [25] to directly estimate the covariance, or Gaussian Graphical Models (GGM) [46] to enforce sparsity on the inverse of the covariance estimate. There have been extensions of the GGM to handle multinomial observations using copulas [39] that have been applied to microbiome analysis [57], [58], [79]. There is also an ongoing effort to extend the GGM to the multiple datasets settings where there is an assumed common precision matrix across the datasets [17], [24], [28], [37], [44], [51], [81], [82]. Notably, [77] extends previous optimization-based approaches to the hierarchical Bayesian setting along with a scalable and efficient inference method. However, this line of work assumes a common feature space across multiple datasets, whereas in our case the features are distinct for different microbial communities.

In the field of computational ecology, there has been a related line of work on joint species distribution modeling (JSDM) [54], [56], [73] to model multiple related abundance datasets. The proposed work differs from JSDM mainly in how we represent the environmental covariates. In JSDM, the environmental covariates are used to infer the species abundance through the generalized linear model, whereas in this work we explicitly represent the covariates through latent variables. With the latter more suitable for applications where the covariates are discrete descriptors of environments such as the binary case (the presence of a bacterium that produces koreenceine antibiotics) we are considering in Section III-B.

Inference in latent variable models, like the one we propose here, can be challenging. This is especially difficult when there is a 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 [16], but this is prone to over-fitting [74] and convergence issues. Another approach is to use Markov Chain Monte Carlo (MCMC) methods, which can be computationally expensive [50], especially in high dimensions. As an alternative, variational Bayes inference has shown much promise [9]. Note that Variational Bayes is not a general purpose method and must be tailored to the specific statistical model [34]. 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 [12]. Additionally, when there is a problematic expression in the joint density, such as the 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 [8] or quadratic [11], [13], [27], [30], [31] functions. We adopt such a local variational bound approach for deriving an inference algorithm for our proposed model.

The proposed model has connections with multi-view learning, text embedding methods, and manifold learning. Supervised PCA [6], [80], Partial Least Squares [66], Canonical Correlation Analysis [32], and Multimodal Factor Analysis (MMFA) [78] allow fusing multi-view data into a common low-dimensional latent space. Among them, only MMFA is applicable to non-Gaussian observations, which, however, does not apply to vectors of count data with observed covariance. Furthermore, the MMFA assumes a non-random latent space, which is known to be prone to over-fitting [74]. Variational auto-encoder-based deep neural network models [71], [76] are often implemented with only a single latent variable to explain multiple modalities. Such autoencoders are implemented by maximizing evidence lower bound (ELBO) that exploits the product of experts framework to combine multiple modalities. [64], [72] use an equalized mixture of experts to combine modality-specific encoder predictions. [69] separates the latent variables as joint and individual where joint latent variables are common for each input modality and individual latent variables are only used to generate the corresponding observations.
Deep generative models have shown recent promise for modeling densities of complex structured data, providing accurate predictions for out-of-sample inputs when the number of training samples is large. However, most microbiome datasets, which are the focus of this paper, have few samples, often many fewer than the number of features. Thus deep models are prone to overfitting such datasets. Furthermore, unlike the proposed model, there is no straightforward way to predict the covariance structure of the observation space using deep learning models.

Count vector data also arises in natural language processing (NLP), where a sentence or a document can be described using a bag-of-words representation. Early NLP models, such as Latent Semantic Analysis/Indexing [19], perform factorization of the count matrix using Singular Value Decomposition, but do not account for the multinomial nature of the data. More recent algorithms, such as Word2Vec [47] and Glove [48], model the sequence of words using a context window. Contemporary representations using methods such as Multidimensional Variable parameterization, it differs significantly from manifold biome assays are not ordered making NLP inapplicable.

Between consecutive words, respectively. Note that, many NLP Networks and Transformers, to model the hidden dynamics are among the most popular, which exploit Recurrent Neural algorithms, such as Word2Vec [47] and Glove [48], model.

bacterial microbiome dataset1.

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

1The code and the dataset are available at https://github.com/maktukmak/microbiome-thor.

A. Notation

We denote the ith count vector replicate for the lth species as \( x_{kl,i} \in \mathbb{Z}^{d_l}_+ \), where \( k \) indexes the experimental condition, and \( d_l \) denotes the total number of transcripts for species \( l \). The total number of experimental conditions 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 \( l = 1, \ldots , L \) and \( k = 1, \ldots , K \). 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 \( I_k \). Concisely, the dataset for experimental condition \( k \) is \( D_k = \{ \{ x_{kl,i} \} | 1 \leq l \leq L \} \).

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 \( \mu_k \) and covariance matrix \( \Sigma_k \). 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 \( z_{k,i} \in \mathbb{R}^{d_k} \) be the latent variable assigned for the data sample \( D_{k,i} \). \( z_{k,i} \) thus has the following multivariate normal prior distribution:

\[
p(z_{k,i}; \mu_k; \Sigma_k) = \mathcal{N}(z_{k,i}; \mu_k, \Sigma_k),
\]

where \( \mu_k \in \mathbb{R}^{d_k} \) is the prior mean vector and \( \Sigma_k \in \mathbb{S}^{d_k}_{++} \) is the positive definite prior covariance matrix. The observed data consists of count vectors of the transcriptomes, which are modeled as multinomial distributed [52]. We model the conditional distributions of the observed count vectors of species \( l \) as follows:

\[
p(x_{kl,i} | z_{k,i}) = Mu(x_{kl,i}; N_{kl,i}, S(\Theta_{kl}z_{k,i})),
\]

where \( N_{kl,i} \) is the total number of counts of the ith data sample of species \( l \) and \( Mu \) denotes the multinomial distribution with the form, \( Mu(x; N, p) = \frac{N!}{\prod_{d=1}^{D} p_d x_d!} \prod_{d=1}^{D} p_d x_d \). Note that we have introduced one more model parameter for each species, specifically \( \Theta_{kl} \in \mathbb{R}^{d_l \times d_k} \), that maps lower dimensional latent space to the higher dimensional observation space of species \( l \). Also note that both the latent variable \( z_{k,i} \) and the parameter \( \Theta_{kl} \) are real-valued. Therefore, to provide a proper simplex support set for the multinomial distribution, we use the softmax function, \( S(\eta)_d = \exp \eta_d / \sum_{d=1}^{D} \exp \eta_d \), where \( S(\eta)_d \) is the \( d \)th element of probability vector \( S(\eta) \) and \( \eta = \Theta_{kl}z_{k,i} \) for notational simplicity. The output of this function is a proper probability vector, i.e., \( \sum_{d=1}^{D} S(\eta)_d = 1 \) and \( S(\eta)_d \geq 0 \) for all \( d = 1, \ldots , D \). See Fig. 1 for a graphical representation of the proposed model.

The lower dimension of the latent variables is a key feature of our model since it explicitly induces lower rank constraints on the observation covariance matrix, as will be explained in Section II-D, leading to a reduction in the total number of model parameters. It also improves the computational efficiency of the optimization algorithms, as shown in Section II-E. A

II. PROPOSED MODEL

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

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
theoretical justification is supplied by the manifold hypothesis [21], which holds that most naturally occurring signals lie in a lower dimensional space, in addition to the principle of Occam’s razor [23], which holds that choosing less complex models leads to better and more stable performance.

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 [8], 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 that 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 [8], [65], categorical PCA [31], and Gaussian process classification [30]. 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

$$\sum_{k=1}^{K} \sum_{i=1}^{l_k} \log p(z_{k,i}, D_{k,i}),$$

in which lse denotes the log-sum-exp function, i.e.,

$$\text{lse}(\eta) = \log \sum_{d=1}^{D} \exp \eta_d,$$

and we suppress the deterministic parameters to avoid clutter. Taking the expectation with respect to $z_{k,i}$ 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], [12] to infer the deterministic parameters $\mu_k$, $\Sigma_k$, and $\Theta_{k,l}$. The main objective is to maximize the likelihood of the observations under the model. The algorithm comprises two alternating steps: i) the Expectation step (E-step), where we integrate out the latent variables, ii) the 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 the 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.C.2 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 $z_{k,i}$ as follows:

$$\text{KL}(q(z_{k,i}) \mid p) = \mathbb{E}_{q(z_{k,i})} \log \frac{q(z_{k,i}; \lambda_{k,i})}{p(z_{k,i} \mid D_{k,i})}$$

$$= \mathbb{E}_{q(z_{k,i})} \log \frac{q(z_{k,i}; \lambda_{k,i})}{p(z_{k,i} \mid D_{k,i})}$$

$$= \mathbb{E}_{q(z_{k,i})} \log \frac{q(z_{k,i}; \lambda_{k,i})}{p(z_{k,i} \mid D_{k,i})} - \log p(z_{k,i}, D_{k,i})$$

(4)

where $\lambda_{k,i}$ corresponds to the set of parameters of the approximate posterior distribution $q(z_{k,i}; \lambda_{k,i})$. The expectation operator is defined as $\mathbb{E}_{q(z)} f(z) = \int f(z) q(z; \lambda) dz$. Note that the evidence (marginal likelihood) $p(D_{k,i})$ does not depend on $z_{k,i}$. 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:

$$\mathcal{L} = \sum_{i=1}^{I} \sum_{k=1}^{K} \mathbb{E}_{q(z_{k,i})} \left[ \log p(z_{k,i}, D_{k,i}) - \log q(z_{k,i}; \lambda_{k,i}) \right],$$

(5)

where the first term in the expectation corresponds to the joint distribution of the latent variable $z_{k,i}$ and the associated observed data $D_{k,i}$. The second term corresponds to the log of the

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
approximate posterior distribution. The joint distribution has the following form:

$$ \log p(z_{k,i}, D_{k,i}) = \log p(z_{k,i}) + \sum_{l=1}^{L} \log p(x_{kl,i}|z_{k,i}). \quad (6) $$

The expressions for $p(x_{kl,i}|z_{k,i})$ and $p(x_{kl,i}|z_{k,i})$ are given in Eqs. (1) and (2), respectively. We approximate the posterior distribution of $z_{k,i}$ as Gaussian with the following form:

$$ q(z_{k,i}; \lambda_{k,i}) = N(z_{k,i}; m_{k,i}, S_{k,i}), \quad (7) $$

where $\lambda_{k,i} = \{m_{k,i}, S_{k,i}\}$ is the set of free parameters. Specifically, $m_{k,i}$ is the posterior mean and $S_{k,i}$ 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 is:

$$ \log p(x_{kl,i}|z_{k,i}) = \sum_{d=1}^{D} x_{kl,id} (\Theta_{kl,d} z_{k,i} - \text{lse}(\Theta_{kl} z_{k,i})), \quad (8) $$

Taking expectation corresponds to integrating out Gaussian distributed $z_{k,i}$. The conditional expectation of the first term is easily determined since it linearly depends on $z_{k,i}$. However, the expectation of the second term, which requires integrating $z_{k,i}$ over the lse function, is intractable to compute in a closed form. To overcome this issue, we perform quadratic surrogate optimization transfer (see Appendix B), in which a quadratic approximation to the lse function [11] 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 (see Appendix C for more details):

$$ \text{lse}(\Theta_{kl} z_{k,i}) \leq \frac{1}{2} z_{k,i}^{T} \Theta_{kl}^{T} A_{l} \Theta_{kl} z_{k,i} - b_{kl,i}^{T} \Theta_{kl} z_{k,i} + c_{kl,i}, \quad (9) $$

where

$$ A_{l} = 0.5[I_{D_{xl}} - (1/(D_{xl} + 1))1_{D_{xl}}1_{D_{xl}}^{T}] \quad (10) $$

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

$$ b_{kl,i} = A_{l} \Phi_{kl,i} - S(\Phi_{kl,i}), \quad (11) $$

$$ c_{kl,i} = \frac{1}{2} \Phi_{kl,i}^{T} A_{l} \Phi_{kl,i} - S(\Phi_{kl,i})^{T} \Phi_{kl,i} + \text{lse}(\Phi_{kl,i}), \quad (12) $$

where $\Phi_{kl,i}$ is the Taylor series expansion point, which is optimized as a free variational parameter. Note that intermediate parameters are a deterministic function of $\Phi_{kl,i}$. Plugging the approximation in Eqs. (5)–(9) results in a convex lower bound on ELBO, denoted as $L'$, which is $\leq L$ and tight at $\Phi_{kl,i}$. Using $L'$ 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 $z_{k,i}$, which are the mean vector $m_{k,i}$ and the covariance matrix $S_{k,i}$. It is straightforward to maximize over these parameters by using the completing-the-square approach [7] (see Appendix A). The terms that quadratically depend on $z_{k,i}$ in the joint log-likelihood yield the posterior covariance update:

$$ S_{k,i} = \left[ \Sigma_{k}^{-1} + \sum_{l=1}^{L} N_{kl,i} \Theta_{kl}^{T} A_{l} \Theta_{kl} \right]^{-1}, \quad (13) $$

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

$$ m_{k,i} = S_{k,i} \left[ \Sigma_{k}^{-1} \mu_{k} + \sum_{l=1}^{L} (x_{kl,i} + N_{kl,i} b_{kl,i}) \Theta_{kl} \right]. \quad (14) $$

Lastly, we update the Taylor series expansion point as:

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

Note that the update of $\Phi_{kl,i}$ depends on the posterior mean. Hence, the algorithm repeats the updates in Eqs. (13)–(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 $L'$ by taking the expectations with respect to the posterior distributions. Afterward, taking the derivatives with respect to the model parameters yields closed-form update equations for the model parameters. Specifically, the updates for each $\Theta_{kl}$ are given as follows:

$$ \Theta_{kl} = \left[ \sum_{i=1}^{I} A_{l}^{-1} (x_{kl,i} + N_{kl,i} b_{kl,i}) m_{k,i}^{T} \right]^{-1} \left[ \sum_{i=1}^{I} N_{kl,i} (m_{k,i} m_{k,i}^{T} + S_{k,i}) \right], \quad (16) $$

where $A_{l}^{-1} = 2[I_{D_{xl}} + ((D_{xl} + 1)/(D_{xl} + 2))1_{D_{xl}}1_{D_{xl}}^{T}]$ using Matrix inversion lemma. The update equations for the mean parameter and covariance of the prior distribution of $z_{k,i}$ then follow as:

$$ \mu_{k} = \frac{1}{I_{k}} \sum_{i=1}^{I} m_{k,i}, \quad (17) $$

$$ \Sigma_{k} = \frac{1}{I} \sum_{i=1}^{I} (m_{k,i} - \mu_{k})(m_{k,i} - \mu_{k})^{T} + S_{k,i}, \quad (18) $$

respectively. Derivations are given in Appendix D and the variational EM algorithm is summarized in Algorithm 1.
enforcing latent variables ambiguous. This problem can be addressed by choosing the maximal converged value. However, we didn’t observe significant improvement in covariance prediction accuracy in the M-step is for computation in the M-step is for computing total complexity of \(O(Ld^3)\) where the covariance matrix \(C_{kl, \text{intra}} = A_l^{-1} + \hat{\Theta}_k \hat{\Sigma}_k \hat{\Sigma}_k^T\) and the mean vector \(\hat{\phi}_k = \Theta_k \mu_k\) is of interest to us, in which \(C_{kl, \text{intra}}\) captures intra-species correlations of species \(l\) in condition \(k\). To obtain inter-species correlations, define \(\hat{A}^{-1} = \text{diag}(A_1^{-1}, \ldots, A_L^{-1})\) and \(\hat{\Theta}_k = [\Theta_{k1}, \ldots, \Theta_{kL}]\), then \(C_{\text{inter}} = \hat{A}^{-1} + \hat{\Theta}_k \hat{\Sigma}_k \hat{\Sigma}_k^T\) 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 \(\text{Corr} = \text{diag}(C)^{-1/2} C \text{diag}(C)^{-1/2}\).

### 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 \(O(Id_z^2)\). Hence, the overall asymptotic complexity for the posterior covariance computation is \(O(I(d_z^2 + Ld_z^2))\). 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 \(O(I(Ld_z d_l + d_z^2 + d_z^2 + Ld_z^2))\). Removing non-dominant terms results in \(O(I(ds_l^2 + Ld_z^2))\). 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 \(\Theta_k\). Eq. (16) comprises two terms. The first term requires \(O(Id_z d_z)\) due to \(I\) times vector-vector outer products. The second term requires \(O(ds_l^2 + d_z^2)\) due to vector-vector outer products and subsequently matrix inversion. Multiplying these terms costs \(O(ds_l^2)\), hence resulting total complexity of \(O(L(Id_z d_z + Id_z^2 + d_z^2 + d_z d_l^2))\) 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
The rank also induces linear complexity in terms of $K$. In summary, both E and M steps scale linearly in terms of $K$, $L$, $d_z$, 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 using only the training dataset. The BIC arises from the Laplace approximation to the model posterior $p(M|D_k)$ [33], where $M$ is the complete model including the latent dimension $d_z$. This results in a Bayesian estimate of $d_z$: $d_z = \arg\max_{d_z} \text{BIC}$, where $\text{BIC} = \log p(D_k) - 0.5 \times \log I_k \times \text{dof}$, which is a function of the total number of unknown parameters that penalizes the log-likelihood with a model complexity penalty term. In the proposed model, we use ELBO lower bound to the likelihood by following [4]. The unknown parameters of the model are $\{\Theta_k\}_{k=1}^K$, $\mu_k$, and $\Sigma_k$. Hence, the total number of parameters is $\text{dof} = K \times (d_z + K)$. 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 50 times to report the performance. The panel on top of Fig. 2 shows the average BIC values obtained after convergence of the variational EM algorithm. We see that maximum BIC values are obtained in the vicinities 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 errors are achieved at the ground truth $d_z$ values, which validates the model selection method.

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 a fixed total number of counts, which is 100, per observation, then stack all the observations. Fig. 3(a) shows the true embeddings of the resulting dataset. We trained the proposed algorithm with the true latent space dimension. Fig. 3(b) shows the embeddings of the model, which are obtained through the posterior distributions. Due to the 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. 3(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. 3(a) and 3(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 $x_{kl,i}$ is an observation-specific parameter, which affects the accuracy of the proposed algorithm. Fig. 4 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 a fixed mean. We also fix the observation dimension to 128. RMSE is reported based on averaging 50 experiments. Fig. 4 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 the low number of counts, the covariance matrix becomes harder to predict due to higher vulnerability to over-fitting. Hence, the lower the dimension of the latent space, the more sensitivity to the total number of counts.

On the other hand, Fig. 5 demonstrates the opposite trend when the dimension of the observation space dimension is increased. Here the total mean counts are fixed at 1000. 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. 4.

4) Baseline Algorithms: Here we present the performance comparisons of the proposed method relative to several baseline methods for estimating the underlying covariance and inverse covariance matrices. i) Empirical covariance, which is computed as the sample covariance. ii) The Ledoit-Wolf estimator [36], which uses shrinkage regularization to perform MAP estimation for the covariance matrix by assigning an inverse Wishart prior to the covariance matrix. iii) Gaussian Copula GraphicalLasso [39], which penalizes the precision matrix with L1-norm constraints after transforming the data by using Gaussian copulas. Regularization forces the entries of the precision matrix to be sparse. iv) Factor Analysis [52] uses another form of regularization of the covariance matrix by imposing a low-rank structure. v) GemBag [77], assumes a common sparsity structure among the environmental conditions by modeling the edges of the environment-specific precision matrices using hierarchical priors. Each of these aforementioned baseline methods is expected to perform best when the data, or its transformed version, is normally distributed. vi) jSDM [73], which is another latent variable model that uses the log link function for the count observations. Environmental conditions and latent variables are graphically joined at the mean vectors, i.e., logits. For the Empirical Covariance, Ledoit-Wolf, GraphicalLasso, FA, and GemBag, we first normalize the data by subtracting the mean and dividing by the variance, before running these methods. On the other hand, the non-Gaussian counting nature of the data is explicitly modeled in our proposed model and jSDM, thus running on the raw observations. For model selection in FA, jSDM, and the proposed model, we use the exact rank of the simulated dataset. The regularization coefficient of the 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 [36]. See Section III in the supplementary material for further implementation details.

5) Simulating Model Communities: Next, we generate a synthetic dataset that 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 $z_i$ with dimension $d_z = 3$ are generated for each measurement site by sampling from $z_i \sim N(0_{d_z}, I_{d_z})$, where $i$ indexes the replicate for $i = 1, \ldots, I$. 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 $\Theta_k \in \mathbb{R}^{d_l \times d_z}$, are chosen randomly by sampling from a zero mean multivariate normal distribution. Then, we sample the observed data $x_{l,i}$ 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 $I = 100$ replicates for each environmental condition where the total number of conditions $K = 2$. The true covariance matrix is then given as $\tilde{\Theta}_k = [\Theta_{k,1}, \Theta_{k,2}]$.

6) Correlation Results: Fig. 6 shows the estimated covariance matrices of the baseline algorithms, the proposed algorithm, alongside the ground truth matrix, when the simulated datasets are realized following Section III.A.5. The proposed model can recover the covariance structure accurately. The relatively poorer accuracy of the other methods can be attributed 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 a low-rank structure, which is not taken into account by the Gaussian Copula Graphical-Lasso, Ledoit-Wolf, or standard sample covariance estimation methods. Third is the common structure assumption of GemBag and jSDM among the covariance/precision matrices for each environment. As the data were simulated from the proposed model, the proposed algorithm naturally performs better. Section IV in the supplementary material discusses more on model mismatch concept with additional simulations. 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 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 [26]. Microlaer model communities are useful to understand principles that govern community behaviors [5], [15], [18], [75]. 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 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.
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 the growth of *F. johnsoniae* 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 the 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 the supplementary material 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 orthologous groupings of gene transcripts into metabolic pathways. Specifically, after pathway mapping, each feature corresponds to a transcriptional 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.A.1. 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 the removal of koreenceine (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 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. Fig. 7 illustrates the effect of the 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 material).

---

2The 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.
especially in *F. johnsoniae* suggests that the removal of koreenceine is associated with an increase in network connectivity (many more vertices whose degrees increase than decrease), orthology IDs with significant changes in vertex mean also have changes in vertex degree, but not conversely. Furthermore, the asymmetry of the blue curve shows the vertex mean 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 wild type. 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 material.

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 [26] 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 Gaussian 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 the exploration of correlation and associated networks for dichotomous microbiome data.

There are several promising directions for future work. One possible area is to generalize the model to capture covariance structures of absence-presence datasets by modeling the binary observations using Bernoulli distributions. Another generalization can be achieved by the incorporation of covariates such as temperature, pH, and physical/chemical perturbations, that may change the composition of the species. The mean of the latent variables can be made a function of the covariates to accomplish that. One another possible area is to incorporate system dynamics into the latent space so as to explicitly capture temporal correlations. In particular, there is increasing interest in collecting longitudinal microbiome data for studying adaptation, resilience, and dynamics over time. The incorporation of a state-space dynamical model into our framework can reveal the 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.

APPENDIX A

ESTIMATION OF POSTERIOR PARAMETERS

Log-likelihood of Multivariate Normal distribution

\[
\log \mathcal{N}(x; \mu, \Sigma) = \frac{1}{2} x^T \Sigma^{-1} x + x^T \Sigma^{-1} \mu + \text{const}
\]

in which the second order term in x corresponds to the inverse of covariance matrix Σ, and the linear term corresponds to the mean when multiplied with Σ. 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 $z_{k,i}$, which is denoted as $q(z_{k,i}; m_{k,i}, S_{k,i})$. Given the joint likelihood in Eq. (6) and quadratic approximation in Eq. (9), one can collect the quadratic terms in $z_{k,i}$ as follows:

$$-\frac{1}{2} z_{k,i}^T \Sigma_k^{-1} z_{k,i} + \sum_{l=1}^{L} \frac{N_{kl,i}}{2} z_{k,i}^T \Theta_{kl}^T A_l \Theta_{kl} z_{k,i},$$

which follows Eq. (13) for posterior covariance $S_{k,i}$ estimate. Similarly, the linear terms are collected as:

$$\Sigma_k^{-1} \mu_k z_{k,i} + \sum_{l=1}^{L} x_{kl,i} \Theta_{kl} z_{k,i} + N_{kl,i} b_{kl,i}^T \Theta_{kl} z_{k,i}.$$ 

Collecting the terms and multiplying with the posterior covariance estimate yields posterior mean $m_{k,i}$ estimate as given in Eq. (14).

**APPENDIX B**

**NOTE ON QUADRATIC SURROGATE OPTIMIZATION TRANSFER**

Due to the non-conjugacy between multinomial and multivariate normal distributions, computing the posterior distributions of the latent variables is intractable, hence we can not obtain closed form expressions for the expectations of the joint likelihood required for the M-step. We adopt an alternative variational inference approach, called quadratic surrogate optimization transfer, where the problematic terms of the joint log-likelihood are replaced with simpler quadratic surrogates obtained by truncated Taylor series expansion. These quadratic functions have tunable free variational parameters and expansion points that control the tightness of the approximation, which are optimized concurrently. This differs from the mean-field approach of variational Bayes inference, which is a global approximation that uses a factorized approximation to the multivariate posterior distribution in order to make the computation of statistical expectation tractable. Quadratic surrogate optimization transfer is on the other hand performed locally for the problematic terms of the joint likelihood, i.e., the approximated quadratic function is created and optimized at each iteration of the EM algorithm. In the literature, this approach has been used for logistic regression [11], multi-task learning [13], discrete factor analysis [30], and correlated topic models [8].

**APPENDIX C**

**UPPER BOUND TO LogSumExp FUNCTION**

For notational simplicity, let $\eta = \Theta_{kl} z_{k,i}$ and drop $k, l, i$ indexes. Second order Taylor series expansion at arbitrary point $\psi$ yields:

$$\text{lse}(\eta) \leq \text{lse}(\psi) + S(\psi)^T (\eta - \psi) + \frac{1}{2}(\eta - \psi)^T A(\eta - \psi),$$

where we replace the original Hessian matrix with constant $A$ in Eq. (10) to obtain the upper bound based on [11]. Reorganizing the terms corresponds to the following quadratic function:

$$\frac{1}{2} \eta^T A \eta - (A \psi - S(\psi))^T \eta + \frac{1}{2} \psi^T A \psi - S(\psi)^T \psi + \text{lse}(\psi).$$

Then, we introduce $b$ and $c$ for linear and constant terms, respectively, to simplify the notation in the main text.

**APPENDIX D**

**DERIVATION OF M-STEP UPDATES**

Taking expectation of $L'$ with respect to the posterior distributions of the latent variables in Eqs. (13) and (14) yields the following expression:

$$\sum_{i=1}^{I_k} \left[ x_{kl,i}^T \Theta_{kl} m_{k,i} - \frac{N_{kl,i}}{2} m_{k,i}^T \Theta_{kl}^T A_l \Theta_{kl} m_{k,i} + \frac{1}{2} \log |\Sigma_k^{-1}| \right]$$

$$- \frac{1}{2}(m_{k,i} - \mu_k)^T \Sigma_k^{-1} (m_{k,i} - \mu_k) - \frac{1}{2} Tr(\Sigma_k^{-1} S_{k,i})$$

$$- \frac{1}{2} N_{kl,i} \text{vec}(\Theta_{kl})^T (A_l \otimes S_{k,i}) \text{vec}(\Theta_{kl}) + \text{const},$$

where vec denotes the vectorization, Tr denotes the trace operator, and $\otimes$ is the Kronecker product. The derivatives of this expression with respect to $\Theta_{kl}$, $\mu_k$, and $\Sigma_k^{-1}$ are given respectively as follows:

$$\Theta_{kl} \rightarrow \sum_{i=1}^{I_k} \left[ x_{kl,i} m_{k,i}^T - \frac{N_{kl,i}}{2} m_{k,i}^T A_l \Theta_{kl} S_{k,i} - N_{kl,i} A_l \Theta_{kl} S_{k,i} \right],$$

$$\mu_k \rightarrow \sum_{i=1}^{I_k} \Sigma_k^{-1} (m_{k,i} - \mu_k),$$

$$\Sigma_k^{-1} \rightarrow \frac{1}{2} \Sigma_k - \frac{1}{2} \sum_{i=1}^{I_k} (m_{k,i} - \mu_k)^T (m_{k,i} - \mu_k) + S_{k,i}.$$

Equating the derivatives to zero results in closed form update equations in Eqs. (16)–(18), for $\Theta_{kl}$, $\mu_k$, and $\Sigma_k$, respectively.

**REFERENCES**

[1] A. Agrawal, A. Ali, and S. Boyd, “Minimum-distortion embedding,” Found. Trends® Mach. Learn., vol. 14, no. 3, pp. 211–378, 2021.

[2] V. Aguiar-Pulido, W. Huang, V. Suarez-Ulloa, T. Cickovski, K. Mathee, and G. Narasimhan, “Metagenomics, metatranscriptomics, and metabolomics approaches for microbiome analysis: Supplementary issue: Bioinformatics methods and applications for big metagenomics data,” Evol. Bioinf. Online, vol. 12, 2016, Art. no. S36436.

[3] C. Archambeau and F. Bach, “Sparse probabilistic projections,” in Proc. Adv. Neural Inf. Process. Syst., vol. 21, pp. 73–80, 2008.

[4] M. J. Beal and Z. Ghahramani, “Variational Bayesian learning of directed graphical models with hidden variables,” Bayesian Anal., vol. 1, no. 4, pp. 793–831, 2006.

[5] J. Bengtsson-Palme, “Microbial model communities: To understand complexity, harness the power of simplicity,” Comput. Struct. Biotech. J., vol. 18, pp. 3987–4001, 2020.

[6] J. Bernardo et al., “Bayesian factor regression models in the ‘large p, small n’ paradigm,” Bayesian Stat., vol. 7, pp. 733–742, 2003.

[7] C. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Vancouver, Canada: Springer, 2006.

[8] D. Blei and J. Lafferty, “Correlated topic models,” in Proc. Adv. Neural Inf. Process. Syst., vol. 18, p. 147, 2006.
B. Jia and F. Liang, “Joint estimation of multiple mixed graphical models,” Journal of Machine Learning Research, vol. 12, no. 1, pp. 515-550, 2001.

T. S. Jaakkola and M. I. Jordan, “Bayesian parameter estimation via variational methods,” Statist. Comput., vol. 14, no. 4, pp. 381-396, 2004.

D. Böhnning, “Multinomial logistic regression algorithm,” Ann. Inst. Stat. Math., vol. 44, no. 1, pp. 197-200, 1992.

L. Bottou, F. E. Curtis, and J. Nocedal, “Optimization methods for large-scale machine learning,” SIAM Rev., vol. 56, no. 2, pp. 223-311, 2014.

M. Braun and J. McAluife, “Variational inference for large-scale models of discrete choice,” J. Am. Stat. Assoc., vol. 105, no. 489, pp. 324-335, 2010.

W. Buntine and A. Jakulin, “Discrete component analysis,” in International Statistical and Optimization Perspectives Workshop “Subspace, Latent Structure and Feature Selection”. Heidelberg, Germany: Springer, 2005, pp. 1-33.

M. G. Chevrette, J. R. Brubtad, C. R. Currie, R. M. Stubbendieck, and J. Guo, E. Levina, G. Michailidis, and J. Zhu, “Joint estimation of multiple graphical models,” J. Multivariate Anal., vol. 10, no. 2, pp. 365-411, 2004.

W. Lee and Y. Liu, “Joint estimation of multiple precision matrices with common structures,” J. Mach. Learn. Res., vol. 16, no. 1, pp. 1035-1062, 2015.

F. Liesecke et al., “Ranking genome-wide correlation measurements improves microarray and RNA-seq based global and targeted co-expression networks,” Science, vol. 398, no. 1, pp. 1-16, 2018.

H. Liu, J. Lafferty, and L. Wasserman, “The nonparanormal: Semiparametric estimation of high dimensional undirected graphs,” J. Mach. Learn. Res., vol. 10, no. 10, pp. 2295-2328, 2009.

Q. Liu and F. A. Kusner, “A survey on contextual embeddings,” 2020, arXiv:2003.07278.

H. F. Lopes and M. West, “Bayesian model assessment in factor analysis,” Stat. Sin., vol. 14, pp. 41-67, 2004.

G. L. Lozano et al., “Introducing THOR, a model microbiome for genetic dissection of community behavior,” mBio, vol. 10, no. 2, 2019, Art. no. e02846-18.

G. L. Lozano et al., “Bacterial analogs of plant tetrahydropyridine alkaloids mediate microbial interactions in a rhizosphere model system,” Appl. Environ. Microbiol., vol. 85, no. 10, 2019, Art. no. e03508-18.

J. Ma and G. Michailidis, “Joint structural estimation of multiple graphical models,” J. Mach. Learn. Res., vol. 17, no. 1, pp. 5777-5824, 2016.

B. Mau and M. A. Newton, “Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo,” J. Comput. Graphical Stat., vol. 6, no. 1, pp. 122-131, 1997.

R. Mazumder and T. Hastie, “The graphical lasso: New insights and alternatives,” Electron. J. Stat., vol. 6, p. 2125, 2012.

T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” 2013, arXiv:1301.3781.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in Proc. Adv. Neural Inf. Process. Syst., vol. 26, pp. 3111-3119, 2013.

G. W. Miller, The Exposeome: A Primer. Amsterdam, The Netherlands: Elsevier, 2013.

S. Mohamed, Z. Ghahramani, and K. A. Heller, “Bayesian exponential family PCA,” in Proc. Adv. Neural Inf. Process. Syst., vol. 21, pp. 1089-1096, 2008.

K. Mohan, P. London, M. Fazel, D. Witten, and S.-I. Lee, “Node-based learning of multiple Gaussian graphical models,” J. Mach. Learn. Res., vol. 15, no. 1, pp. 445-488, 2014.

K. P. Murphy, Machine Learning: A Probabilistic Perspective. Cambridge, MA, USA: MIT Press, 2012.

K. P. Murphy, Probabilistic Machine Learning: Advanced Topics. Cambridge, MA, USA: MIT Press, 2023.

O. Ovaskainen and N. Abrego, Joint Species Distribution Modelling: With Applications in R. Cambridge, U.K.: Cambridge Univ. Press, 2020.

M. E. Peters et al., “Deep contextualized word representations,” Feb. 2018, arXiv:1802.05365.

L. J. Pollock et al., “Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (JSDM),” Methods Ecol. Evol., vol. 5, no. 5, pp. 397-406, 2014.

G. C. Popovic, F. K. Hui, and D. I. Warton, “A general algorithm for covariance modeling of discrete data,” J. Multivariate Anal., vol. 165, no. 86, pp. 86-100, 2018.

G. C. Popovic, D. I. Warton, F. J. Thomson, F. K. Hui, and A. T. Moles, “Untangling direct species associations from indirect mediator species effects with graphical models,” Methods Ecol. Evol., vol. 10, no. 9, pp. 1571-1583, 2019.

M. Rattray, O. Stegle, K. Sharp, and J. Winn, “Inference algorithms and learning theory for Bayesian sparse factor analysis,” J. Phys.: Conf. Ser., vol. 197, 2009, Art. no. 012002.

F. J. Rohlf and R. R. Sokal, “Comparing numerical taxonomic studies,” Syst. Biol., vol. 30, no. 4, pp. 459-490, 1981.

S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” Science, vol. 290, no. 5500, pp. 2323-2326, 2000.

B. Schölkopf, A. Smola, and K.-R. Müller, “Nonlinear component analysis as a kernel eigenvalue problem,” Neural Comput., vol. 10, no. 5, pp. 1299-1319, 1998.
[63] M. Shakya, C.-C. Lo, and P. S. Chain, “Advances and challenges in metatranscriptomic analysis,” *Front. Genet.*, vol. 10, p. 904, 2019.

[64] Y. Shi, N. Siddharth, B. Paige, and P. Torr, “Variational mixture-of-experts autoencoders for multi-modal deep generative models,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 15718–15729.

[65] A. Srivastava and C. Sutton, “Autoencoding variational inference for topic models,” 2017, arXiv:1703.01488.

[66] L. Sun, S. Ji, S. Yu, and J. Ye, “On the equivalence between canonical correlation analysis and orthonormalized partial least squares,” in *Proc. IJCAI*, vol. 9, 2009, pp. 1230–1235.

[67] J. B. Tenenbaum, V. Silva, and J. C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[68] M. E. Tipping and C. M. Bishop, “Probabilistic principal component analysis,” *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)*, vol. 61, no. 3, pp. 611–622, 1999.

[69] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, “Learning factorized multimodal representations,” in *Proc. Int. Conf. Representation Learn.*, 2019, pp. 1–20.

[70] Maaten and G. Hinton, “Visualizing data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[71] R. Vedantam, I. Fischer, J. Huang, and K. Murphy, “Generative models of visually grounded imagination,” 2017, arXiv:1705.10762.

[72] W. Wang, X. Yan, H. Lee, and K. Livescu, “Deep variational canonical correlation analysis,” 2016, arXiv:1610.03454.

[73] D. I. Warton et al., “So many variables: Joint modeling in community ecology,” *Trends Ecol. Evol.*, vol. 30, no. 12, pp. 766–779, 2015.

[74] M. Welling, C. Chemudugunta, and N. Sutter, “Deterministic latent variable models and their pitfalls,” in *Proc. SIAM Int. Conf. Data Min.*, Philadelphia, PA, USA: SIAM, 2008, pp. 196–207.

[75] B. E. Wolfe, “Using cultivated microbial communities to dissect microbiome assembly: Challenges, limitations, and the path ahead,” *mSystems*, vol. 3, no. 2, 2018, Art. no. e00161-17.

[76] M. Wu and N. Goodman, “Multimodal generative models for scalable weakly-supervised learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 5575–5585.

[77] X. Yang, L. Gan, N. N. Narisetty, and F. Liang, “Gembag: Group estimation of multiple Bayesian graphical models,” *J. Mach. Learn. Res.*, vol. 22, 2021, Art. no. 54.

[78] Y. Yilmaz, M. Aktukmak, and A. O. Hero, “Multimodal data fusion in high-dimensional heterogeneous datasets via generative models,” *IEEE Trans. Signal Process.*, vol. 69, pp. 5175–5188, 2021.

[79] G. Yoon, I. Gaynanova, and C. L. Müller, “Microbial networks in SPRING—Semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data,” *Front. Genet.*, vol. 10, p. 516, 2019.

[80] S. Yu, K. Yu, V. Tresp, H.-P. Kriegel, and M. Wu, “Supervised probabilistic principal component analysis,” in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Min.*, 2006, pp. 464–473.

[81] B. Zhang and Y. Wang, “Learning structural changes of Gaussian graphical models in controlled experiments,” 2012, arXiv:1203.3532.

[82] Y. Zhu and R. F. Barber, “The log-shift penalty for adaptive estimation of multiple Gaussian graphical models,” in *Proc. Artif. Intell. Stat.*, San Diego, CA, USA, PMLR, 2015, pp. 1153–1161.