Statistics in Biosciences (2021) 13:351–372
https://doi.org/10.1007/s12561-020-09294-z

Joint Microbial and Metabolomic Network Estimation with the Censored Gaussian Graphical Model

Jing Ma[1]

Received: 15 March 2019 / Revised: 3 September 2020 / Accepted: 8 September 2020 /
Published online: 21 September 2020

Abstract
Joint analysis of microbiome and metabolomic data represents an imperative objective as the field moves beyond basic microbiome association studies and turns towards mechanistic and translational investigations. We present a censored Gaussian graphical model framework, where the metabolomic data are treated as continuous and the microbiome data as censored at zero, to identify direct interactions (defined as conditional dependence relationships) between microbial species and metabolites. Simulated examples show that our method metaMint performs favorably compared to the existing ones. metaMint also provides interpretable microbe-metabolite interactions when applied to a bacterial vaginosis data set. R implementation of metaMint is available on GitHub.

Keywords Data integration · Microbiome · Metabolomics · Censored Gaussian graphical models · Conditional dependence

1 Introduction

The field of microbiome research is shifting rapidly from cataloging the taxonomic compositions of microbial communities [1] to refined technologies that capture strain-level variations or amplicon sequence variants [2–4] and to multi-omics studies that better capture community functional activity [5]. In particular, metabolomics has been extremely useful in explaining microbial functional potential because of its capability in tracking microbially derived metabolites [6–8]. Associations between specific microbes and metabolites provide key insights and improved mechanistic models of host-microbe interactions [9–12]. In practice, the non-parametric Spearman's rank correlation is often used to quantify the pairwise correlation between microbes and metabolites. However, Spearman's rank correlation only captures marginal monotonic

✉ Jing Ma
jingma@fredhutch.org

[1] Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

association and does not distinguish direct and indirect interactions. In contrast, partial correlations measure conditional dependencies and allow the identification of direct interactions between microbes and metabolites [13].

One analytical challenge specific to the microbiome data are the uneven sequencing depths that arise due to differential efficiency of the sequencing process. The total number of reads in a sample is also constrained by the biological specimen at hand and does not reflect the absolute abundance present in the ecosystem. A common practice to address this issue is to transform the raw counts into relative abundances by normalizing over the total sequencing reads in each sample. In other words, raw sequencing counts are transformed into proportions of different microbes whose sum has to be one, also known as compositional data. Several lines of work have been proposed to model marginal and/or conditional microbial interactions from compositional data. For example, SparCC [14] and CCLasso [15] both estimate the linear Pearson correlations between log-transformed counts. A major limitation of marginal association measures such as the Pearson correlation is that they cannot distinguish between direct and indirect relationships [16]. To address this issue, SPIEC-EASI [17] learns the conditional dependencies between pairs of microbes while adjusting for effects from other species in the analysis. This is achieved by estimating the inverse covariance of the centered log-ratio (clr) transformed data using e.g., the graphical lasso algorithm [18]. Fang et al. [19] assume that the observed relative abundances follow the logistic normal distribution and proposed a Majorization-Minimization algorithm for learning the conditional dependence relationships among microbes.

Many of the aforementioned methods are specific to microbiome data and are not directly applicable for joint analysis of microbiome and other omics data types. One naive approach for joint estimation is to apply the graphical lasso algorithm directly to clr transformed microbiome and metabolomic data. However, as illustrated in Fig. 1, the Gaussian graphical model may be a poor fit for microbiome data because the marginal distributions of transformed raw counts are in fact highly skewed and often zero inflated.

This motivates the need for new statistical methodology that can accommodate both microbiome and metabolomic data while accounting for the zero inflation in microbial abundance. Some zero values are sampling zeros that arise due to limited sequencing depths, whereas others are biological zeros that indicate complete absence of a species [20]. Silverman et al. [21] in an unpublished manuscript illustrated that biological zeros in many applications can be approximated as sampling zeros because they both represent a truly low abundance. In this paper, we treat the observed zeros as due to undersampling, and propose a censored Gaussian graphical model (cGGM) to infer the conditional dependencies among microbes and metabolites. Specifically, let $\boldsymbol{W} = (W_1, \ldots, W_q)^\intercal$ with $W_j > 0$ for all j be the latent variables, called *the basis*, that represent the true absolute abundance for each species. Due to undersampling and uneven sequencing depths, the observed abundance $\boldsymbol{R}$ is related to $\boldsymbol{W}$ via

$$R_j = NW_j \boldsymbol{I}(\log W_j > u_j), \tag{1}$$

where $N > 0$ is a scaling factor that may depend on $\boldsymbol{W}$, u_j is a constant which indicates the limit of detection for the j-th variable, and $\boldsymbol{I}(\cdot)$ is the indicator function.

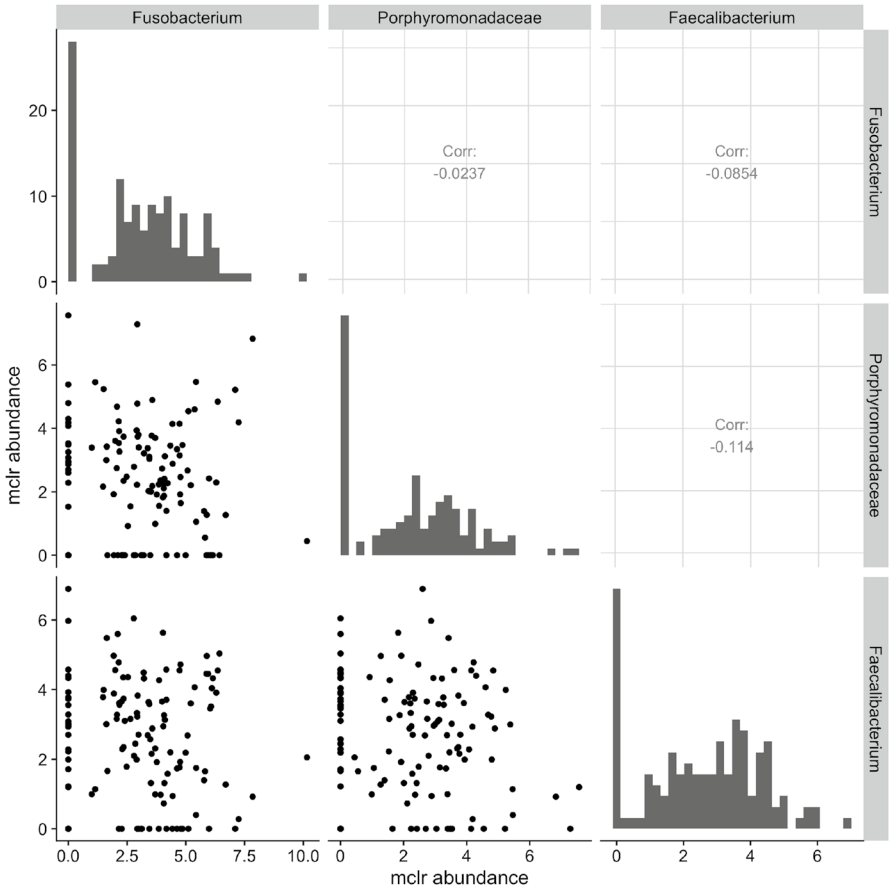

Fig. 1 Scatter plots of the modified centered log-ratio (mclr) transformed abundances of 3 bacterial species in the vaginal microbiome data from McMillan et al. [9]. Marginal distribution of each species is illustrated along the diagonal. The upper panels show the Pearson correlations between pairs of species

The censoring value u_j may be known from the experiment or estimated from data. To adjust for the uneven sequencing depths, we apply the modified clr (mclr) transformation to $\boldsymbol{R}$, which transforms all non-zero counts using the usual clr and shifts all transformed values to be strictly positive [22]. The diagonal panels in Fig. 1 show the histograms of mclr transformed abundances. Compared to the usual clr transformation that requires a pseudo count when dealing with zeros, mclr preserves the ranking of observed counts across multiple samples and is less biased towards rare species [22]. Denote $\boldsymbol{X}_1 = \text{mclr}_\epsilon(\boldsymbol{R})$ the resulting vector after mclr transformation with parameter ϵ, which we elaborate in Sect. 2.3. Let $\boldsymbol{X}_2 = (X_{q+1}, \ldots, X_p)^\intercal$ denote the log transformed concentration measures from $p - q$ $(p > q)$ metabolites. A natural model for integrating microbiome and metabolomic data is to assume that $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ follow a censored multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance

Σ. Zero entries in the inverse covariance matrix $\Omega = \Sigma^{-1}$ capture the conditional independence relationships among the microbes and metabolites.

The problem of inferring the joint microbe-metabolite network thus reduces to estimating Ω from n independent and identically distributed observations on $(\boldsymbol{X}_1, \boldsymbol{X}_2)$. We provide metaMint which is based on estimating each pair of marginal correlations with maximum likelihood. Given the estimated correlation matrix, metaMint uses the graphical lasso to recover the conditional dependencies between microbes and metabolites (direct interactions). We compare our method with several existing approaches in simulations, and show that metaMint outperforms the others in network structural recovery and accuracy of estimating the inverse covariance matrix. When applied to a real data on bacterial vaginosis [9], the integrated network reveals biologically relevant microbe-metabolite interactions and also identifies novel interactions that may serve as potential biomarkers for diagnosis and treatment of bacterial vaginosis.

The censored multivariate normal distribution has been commonly used to analyze environmental data that are often subject to pre-specified detection limits. For example, Hoffman and Johnson [23], Pesonen et al. [24] and Jones et al. [25] studied covariance estimation for left censored multivariate normal distribution in the classic low-dimensional setting. Recently, Augugliaro et al. [26] proposed an approximated EM algorithm for inverse covariance estimation in the high-dimensional setting and applied the method to single-cell data. The work by McDavid et al. [27] was also motivated by single-cell data, but the authors proposed the zero-inflated Gaussian graphical model, which treats zeros as coming from a degenerate point mass at zero instead of being censored. Compared to existing literature, our contribution is a unified model for joint estimation of the integrated microbe and metabolite network in the high-dimensional setting. Our algorithm works well in a variety of scenarios.

The rest of the paper is organized as follows. In Sect. 2, we describe the censored Gaussian graphical model framework and the proposed algorithm. We present extensive numerical studies in Sect. 3 and a real data example on bacterial vaginosis in Sect. 4. We conclude our paper with discussions in Sect. 5.

2 The Censored Gaussian Graphical Model

The censored Gaussian graphical model is suitable for zero-inflated data, which is often the case with microbiome data as shown in Fig. 1. In practice, it is reasonable to assume that the observed zeros are due to undersampling or censoring from below.

Definition 1 A random vector $\boldsymbol{X}$ is said to follow a censored multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance Σ if there exists constants $u_1, \ldots, u_p$ such that $X_j = Y_j \boldsymbol{I}(Y_j > u_j) + u_j \boldsymbol{I}(Y_j \leq u_j)$ where

$$\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \Sigma).$$

The censoring values $\boldsymbol{u} = (u_1, \ldots, u_p)^\intercal$ are experiment specific and can be inferred from data. For example, one can use the smallest value that occurs more than a pre-specified threshold (e.g. 10%) as an estimate. A pre-specified threshold is necessary to ensure that the smallest value occurs more often than by chance. For zero-inflated microbiome data, the censoring values are set to be 0. When there is no censoring in the j-th variable, we set $u_j = -\infty$.

The density of the multivariate normal distribution with mean $\boldsymbol{\mu}$ and inverse covariance $\Omega = \Sigma^{-1}$ is

$$\phi(\boldsymbol{y};\boldsymbol{\mu}, \Omega) = (2\pi)^{-p/2}|\Omega|^{1/2} \exp\{(\boldsymbol{y} - \boldsymbol{\mu})^\intercal \Omega(\boldsymbol{y} - \boldsymbol{\mu})\}.$$

Without loss of generality, let $\boldsymbol{X} = (\boldsymbol{X}_o, \boldsymbol{X}_c)$ where $\boldsymbol{X}_o$ denotes the uncensored components and $\boldsymbol{X}_c$ denotes the censored components. Given censoring values $\boldsymbol{u} = (-\infty, \ldots, -\infty, \boldsymbol{u}_c)$, the density function of $\boldsymbol{X}$ is

$$\psi(\boldsymbol{x}_o, \boldsymbol{u};\boldsymbol{\mu}, \Omega) = \int_{\boldsymbol{u}_c}^{\infty} \phi(\boldsymbol{x}_o, \boldsymbol{x}_c;\boldsymbol{\mu}, \Omega)\mathrm{d}\boldsymbol{x}_c = \phi(\boldsymbol{x}_o;\boldsymbol{\mu}, \Omega) \int_{\boldsymbol{u}_c}^{\infty} \phi(\boldsymbol{x}_c \mid \boldsymbol{x}_o;\boldsymbol{\mu}, \Omega)\mathrm{d}\boldsymbol{x}_c. \tag{2}$$

Let $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\}$ denote a set of n independent and identically distributed observations on $\boldsymbol{X}$. In high-dimensional settings, a natural strategy to estimate the inverse covariance matrix is to maximize the ℓ_1 penalized loss function

$$\frac{1}{n}\sum_{i=1}^{n} \log \psi(\boldsymbol{x}^{(i)}, \boldsymbol{u};\boldsymbol{\mu}, \Omega) - \lambda_n \sum_{1 \le j < k \le p} |\Omega_{jk}|, \tag{3}$$

where λ_n is a regularization parameter that controls the sparsity of Ω. However, direct optimization of (3) is challenging due to the integral in (2) over a potentially high-dimensional space. Augugliaro et al. [26] studied a general version of (3) where variables can be left and right censored. They proposed to use the EM algorithm to optimize the expectation of the full log-likelihood with respect to the conditional distribution $\boldsymbol{X}_c \mid \boldsymbol{X}_o$. However, exact optimization of the EM algorithm is computationally challenging as it requires the second moment of $\boldsymbol{X}_c \mid \boldsymbol{X}_o$, which is a multivariate truncated Gaussian. The approximation in Augugliaro et al. [26] is adapted from Guo et al. [28] and only works well when the inverse covariance matrix is very sparse or the regularization parameter λ_n is large.

2.1 A Direct Estimator Via Marginal Correlations

Our proposal metaMint is based on estimating the marginal correlations directly. A similar idea was used to estimate the correlation matrix of ordinal graphical models [29], where the authors showed that the direct estimator achieves more accurate estimation of the inverse covariance matrix compared to the approximated EM approach in Guo et al. [28].

The first step in metaMint is to estimate the marginal distribution for each variable, which can be done by fitting a univariate Tobit model [30] and has been implemented in the R package `censReg` [31]. Let $\hat{\mu}_j$ and $\hat{\sigma}_j^2$ be, respectively, the estimate of the

mean and variance for the j-th variable. It can be shown that $\hat{\mu}_j$ is a consistent estimate of μ_j, and $\hat{\sigma}_j^2$ is consistent for $\sigma_j^2 = \Sigma_{jj}$. To find the empirical covariance matrix $\hat{\Sigma}$, it suffices to estimate each pairwise correlation.

Suppose we have two variables X_j and X_k ($j < k$). If no observation is censored, it is straightforward to estimate their correlation using the Pearson's correlation coefficient. In the following, we provide details on correlation estimation when at least one variable is censored.

Consider first the case where both variables X_j and X_k are censored from below with u_j and u_k, respectively. For the i-th observation, let $\eta_{ij} = \boldsymbol{I}(x_j^{(i)} > u_j)$ be the indicator function of whether the j-th variable is censored. The pairwise joint log-likelihood can be written as a function of the correlation ρ_{jk},

$$\begin{aligned}\ell_1^{(i)}(\rho_{jk};\mu_j,\mu_k,\sigma_j^2,\sigma_k^2) = {} & \eta_{ij}\eta_{ik}\log \mathrm{P}(Y_j = x_j^{(i)}, Y_k = x_k^{(i)}) \\ & + \eta_{ij}(1-\eta_{ik})\log \mathrm{P}(Y_j = x_j^{(i)}, Y_k < u_k) \\ & + (1-\eta_{ij})\eta_{ik}\log \mathrm{P}(Y_j < u_j, Y_k = x_k^{(i)}) \\ & + (1-\eta_{ij})(1-\eta_{ik})\log \mathrm{P}(Y_j < u_j, Y_k < u_k),\end{aligned}$$

where Y_j and Y_k are bivariate normal with mean $(\mu_j, \mu_k)^\intercal$ and covariance

$$\begin{pmatrix} \sigma_j^2 & \rho_{jk}\sigma_j\sigma_k \\ \rho_{jk}\sigma_j\sigma_k & \sigma_k^2 \end{pmatrix}.$$

Let $\phi(\cdot)$ and $\boldsymbol{\Phi}(\cdot)$ denote, respectively, the density and the cumulative distribution function (c.d.f.) of a standard normal variable. Let the c.d.f. of a bivariate standard normal variable with correlation ρ be $\boldsymbol{\Phi}_2(u, v, \rho)$. The conditional distribution $Y_k \mid Y_j = x^{(i)}$ is again a normal distribution with mean $\tilde{\mu}_k = \mu_k + \frac{\sigma_k}{\sigma_j}\rho_{jk}(x_j^{(i)} - \mu_j)$ and standard deviation $\tilde{\sigma}_k = \sigma_k\sqrt{1-\rho_{jk}^2}$. The pairwise joint log-likelihood thus becomes

$$\begin{aligned}\ell_1^{(i)}(\rho_{jk};\mu_j,\mu_k,\sigma_j^2,\sigma_k^2) = {} & \eta_{ij}\eta_{ik}\log\left\{\frac{1}{\tilde{\sigma}_k}\phi\left(\frac{x_k^{(i)}-\tilde{\mu}_k}{\tilde{\sigma}_k}\right)\frac{1}{\sigma_j}\phi\left(\frac{x_j^{(i)}-\mu_j}{\sigma_j}\right)\right\} \\ & + \eta_{ij}(1-\eta_{ik})\log\left\{\boldsymbol{\Phi}\left(\frac{u_k-\tilde{\mu}_k}{\tilde{\sigma}_k}\right)\frac{1}{\sigma_j}\phi\left(\frac{x_j^{(i)}-\mu_j}{\sigma_j}\right)\right\} \\ & + (1-\eta_{ij})\eta_{ik}\log\left\{\boldsymbol{\Phi}\left(\frac{u_j-\tilde{\mu}_j}{\tilde{\sigma}_j}\right)\frac{1}{\sigma_k}\phi\left(\frac{x_k^{(i)}-\mu_k}{\sigma_k}\right)\right\} \\ & + (1-\eta_{ij})(1-\eta_{ik})\log\boldsymbol{\Phi}_2\left(\frac{u_j-\mu_j}{\sigma_j}, \frac{u_k-\mu_k}{\sigma_k}, \rho_{jk}\right),\end{aligned}$$

where

$$\tilde{\mu}_j = \mu_j + \frac{\sigma_j}{\sigma_k}\rho_{jk}(x_k^{(i)} - \mu_k), \quad \tilde{\sigma}_j = \sigma_j\sqrt{1-\rho_{jk}^2}.$$

If $u_j = -\infty$, this yields a bivariate random vector with only the first variable being censored. Then the joint log-likelihood becomes

$$\begin{aligned}\ell_2^{(i)}(\rho_{jk};\mu_j,\mu_k,\sigma_j^2,\sigma_k^2) =& \eta_{ik}\log \mathrm{P}(Y_j = x_j^{(i)}, Y_k = x_k^{(i)}) \\ &+ (1-\eta_{ik})\log \mathrm{P}(Y_j = x_j^{(i)}, Y_k < u_k) \\ =& \eta_{ik}\log\left\{\frac{1}{\tilde{\sigma}_k}\phi\left(\frac{x_k^{(i)}-\tilde{\mu}_k}{\tilde{\sigma}_k}\right)\frac{1}{\sigma_j}\phi\left(\frac{x_j^{(i)}-\mu_j}{\sigma_j}\right)\right\} \\ &+ (1-\eta_{ik})\log\left\{\Phi\left(\frac{u_k-\tilde{\mu}_k}{\tilde{\sigma}_k}\right)\frac{1}{\sigma_j}\phi\left(\frac{x_j^{(i)}-\mu_j}{\sigma_j}\right)\right\}.\end{aligned}$$

We can solve for ρ_{jk} as

$$\hat{\rho}_{jk} = \arg\max_{\rho\in(-1,1)} \frac{1}{n}\sum_{i=1}^{n} \ell_h^{(i)}(\rho;\hat{\mu}_j,\hat{\mu}_k,\hat{\sigma}_j^2,\hat{\sigma}_k^2), \quad h = 1,2. \tag{4}$$

Because entries in $\hat{\Sigma}$ are estimated separately, $\hat{\Sigma}$ is not guaranteed to be positive semi-definite, which is unsatisfactory because ideally we expect the empirical covariance matrix to be positive semi-definite. One way of bypassing this issue is to use the projection of $\hat{\Sigma}$ onto a positive semi-definite cone, as done in Fan et al. [32]. In practice, one can calculate the eigen-decomposition of $\hat{\Sigma}$ and threshold the negative ones to zero, which yields a new estimator $\widetilde{\Sigma}$.

Given $\widetilde{\Sigma}$, one can apply the graphical lasso algorithm [18]

$$\widetilde{\Omega} = \arg\max_{\Omega}\left\{\log\det(\Omega) - \mathrm{tr}(\widetilde{\Sigma}\Omega) - \lambda_n\sum_{1\le j<k\le p}|\Omega_{jk}|\right\}, \tag{5}$$

to solve for the inverse covariance matrix Ω.

Remark 1 The graphical lasso in (5) can be replaced with other algorithms for inverse covariance matrix estimation such as the method by Cai et al. [33] or its adaptive version [34].

metaMint has been implemented in R. In particular, the optimization in (4) is solved using the `optim` function in R, and (5) is solved by the graphical lasso algorithm in the `glasso` package.

2.2 Tuning Parameter Selection

As with other penalization-based methods, the proposed algorithm requires the specification of a tuning parameter λ_n that controls the sparsity of the inverse covariance matrix. One can use the cross validation procedure in Guo et al. [28] or the stability approach in Liu et al. [35] to select the optimal parameter. In simulations where the ground truth is known, model selection can also be done by maximizing the accuracy in network structural recovery. In Sect. 4 on real data analysis, we used the stability approach in Liu et al. [35].

2.3 The Modified Centered Log-Ratio

The centered log-ratio transformation is often used to transform observed microbial counts to values that are comparable across samples before downstream analysis [36–38]. Let $g(\boldsymbol{r}) = (\Pi_{j=1}^{p} r_j)^{1/p}$ denote the geometric mean of $\boldsymbol{r} = (r_1, \ldots, r_p)$. The clr of $\boldsymbol{r}$ is defined as

$$\mathrm{clr}(\boldsymbol{r}) = \left(\log \frac{r_1}{g(\boldsymbol{r})}, \ldots, \log \frac{r_p}{g(\boldsymbol{r})} \right)^{\intercal}.$$

In practice, each sample may consist of many rare species that have zero counts. Thus a pseudo count of 0.5 or 1 is often added to all counts before clr is applied. However, this practice may unfairly bias rare species and impact the accuracy in correlation estimation. The modified centered log-ratio (mclr) [22] attempts to address this limitation by transforming the non-zero counts with the usual clr and shifting all transformed values to be strictly positive.

Without loss of generality, let $\boldsymbol{r}^{(i)} = (\boldsymbol{r}_1^{(i)}, \boldsymbol{0})^{\intercal} = (N_i \boldsymbol{w}_1^{(i)}, \boldsymbol{0})^{\intercal}$ where only components in $\boldsymbol{r}_1^{(i)}$ (and $\boldsymbol{w}_1^{(i)}$) are positive. Although the sample-specific scaling factor N_i does not affect the relative abundances in sample i, it captures the variation among total sequencing reads. For example, Vandeputte et al. [39] observed up to tenfold differences in the total microbial loads after correcting for microbial cell counts. We define $\mathrm{mclr}_{\varepsilon}(\boldsymbol{r}^{(i)})$ as $(\mathrm{clr}(\boldsymbol{r}_1^{(i)}) + \varepsilon, \boldsymbol{0})^{\intercal}$, where the constant ε is set to be $|\min_{i,j} \log\{r_j^{(i)}/g(\boldsymbol{r}_1^{(i)}\}| + c$ and $c > 0$ is a small constant used to differentiate small positive counts from observed zeros. The resulting $\mathrm{mclr}_{\varepsilon}(\boldsymbol{r}^{(i)})$ is independent of the scaling factor N_i, because $\mathrm{clr}(\boldsymbol{r}_1^{(i)}) = \mathrm{clr}(\boldsymbol{w}_1^{(i)})$. However, adding a pseudo count to zeros and applying clr may introduce unnecessary bias towards zero counts. Figure 2 illustrates the marginal distributions of the genus *Fusobacterium* after the two transformations. Compared to clr, the mclr preserves the relative ranking of all counts while adjusting for the total sequencing depths.

Lastly, it is worth mentioning that mclr defined above is equivalent to transforming the relative abundances as done in Yoon et al. [22]. To see this, let the relative abundance $\boldsymbol{z}^{(i)}$ be defined such that $z_j^{(i)} = r_j^{(i)}/S$, where $S = \sum_{j=1}^{p} r_j^{(i)}$. Moreover, we can write $\boldsymbol{z}^{(i)} = (\boldsymbol{z}_1^{(i)}, \boldsymbol{0})^{\intercal}$ such that only components in $\boldsymbol{z}_1^{(i)}$ are positive. For any $z_j^{(i)} > 0$,

$$\log \frac{z_j^{(i)}}{g(z_1^{(i)})} = \log \frac{r_j^{(i)}}{S} - \{\log g(\boldsymbol{r}_1^{(i)}) - \log S\} = \log \frac{r_j^{(i)}}{g(\boldsymbol{r}_1^{(i)})}.$$

In other words, mclr is scale invariant.

3 Simulation Studies

3.1 Model Setup

We first generated $\boldsymbol{y}^{(i)}$ $(i = 1, \ldots, n)$ from a multivariate normal distribution with mean $\boldsymbol{\mu}_0$ and inverse covariance Ω_0. The mean parameter $\boldsymbol{\mu}_0$ was generated uniformly from $[-0.5, 2]$ to reflect the heterogeneity in abundances of microbial sequences and metabolites. To generate the inverse covariance matrix Ω_0, we considered the following network models, each with p nodes:

(1) Scale-free network. This network was generated using the Barabasi-Albert algorithm [40] and has $(p - 1)$ edges. The left panel of Fig. 3 illustrates a scale-free network.
(2) Erdős-Rényi random graph [41]. This network has p edges, as illustrated in the middle panel of Fig. 3.
(3) Nearest-neighbor network. We constructed this network using the same procedure described in Guo et al. [28], where we uniformly sampled p points on a unit square and linked any two points that are 5 nearest neighbors of each other in terms of their Euclidean distances. This network has about $2.5p$ edges. The right panel of Fig. 3 illustrates one realization of a sparse network generated with 2 nearest neighbors.

Given the network topology, the off-diagonal entries in Ω_0 were generated uniformly from $[-1, -0.5] \cup [0.5, 1]$, with diagonal entries being $|\Lambda_{\min}(\Omega_0^-)| + 0.1$. Here Ω_0^- represents the matrix Ω_0 with zeros in the diagonal and $\Lambda_{\min}(A)$ denotes the smallest eigenvalue of A. The covariance matrix Σ_0 is then determined by

$$\Sigma_{0,jk} = (\Omega_0)_{jk}^{-1} / \sqrt{(\Omega_0)_{jj}^{-1}(\Omega_0)_{kk}^{-1}}.$$

By construction, the diagonal entries of Σ_0 are all 1.

Given the latent $\boldsymbol{y}^{(i)}$, the basis vector $\boldsymbol{w}^{(i)} = (w_1^{(i)}, \ldots, w_p^{(i)})^\intercal$ was obtained through the transformation $w_j^{(i)} = e^{y_j^{(i)}}$. Censored abundances $\boldsymbol{r}^{(i)} = (r_1^{(i)}, \ldots, r_p^{(i)})^\intercal$ were generated such that

$$r_j^{(i)} = \begin{cases} N_i w_j^{(i)} \boldsymbol{I}(y_j^{(i)} > 0) & j = 1, \ldots, q, \\ w_j^{(i)} & j = q+1, \ldots, p, \end{cases}$$

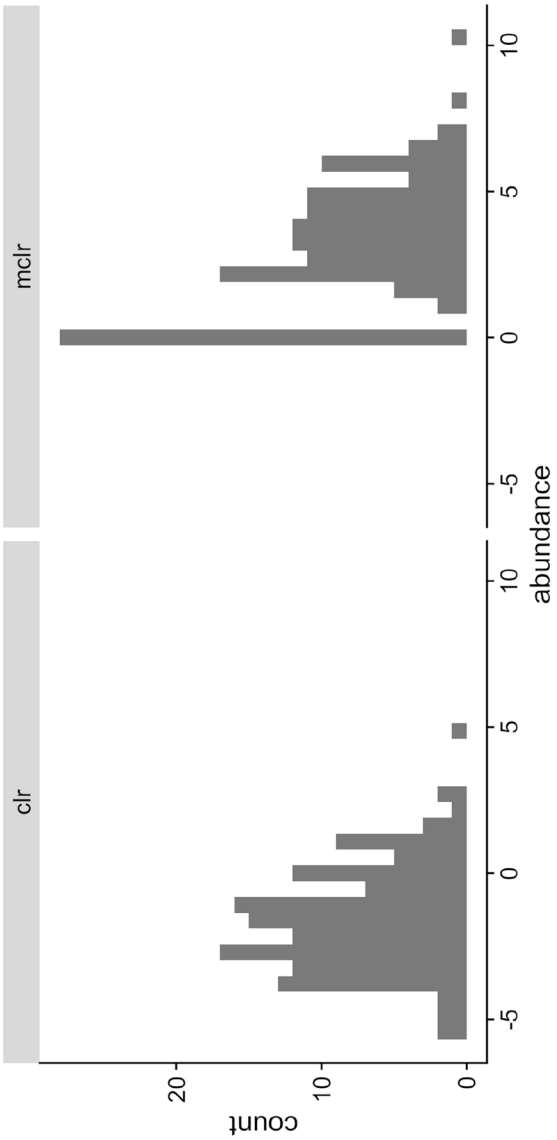

Fig. 2 Marginal distributions of the Fusobacterium genus after the clr (left) and mclr (right) transformation to the 131 observations used in McMillan et al. [9]. A pseudo count of 0.5 was added to all counts in order to apply the clr transformation, whereas $c = 0.1$ for mclr

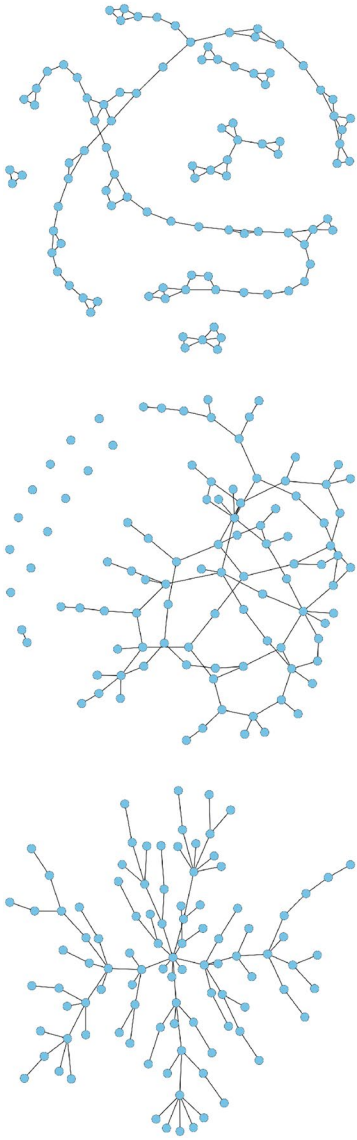

Fig. 3 Illustration of graphs used in our simulations ($p = 100$): scale-free graph (left), random graph (middle), and nearest-neighbor graph (right)

where N_i is generated uniformly between 1 and 10. Here q indicates the number of microbes. Only microbiome data are assumed to be censored and compositional in this article, but this assumption can be relaxed in general. In all simulations, we set the constant $c = 0.1$ in the modified clr transformation. Denote $\boldsymbol{x}_1^{(i)} = \text{mclr}_\epsilon(\boldsymbol{r}_{1:q}^{(i)})$ and the observed abundances $\boldsymbol{x}^{(i)} = (\boldsymbol{x}_1^{(i)}, \log r_{q+1}^{(i)}, \ldots, \log r_p^{(i)})^\intercal$.

3.2 Results

We compared metaMint with SPIEC-EASI [17] and gCoda [19]. The oracle estimator obtained from the latent basis $\{\boldsymbol{w}^{(i)}\}_{i=1}^n$ is used as a benchmark, though in practice the oracle is generally unknown. To evaluate the performance of network recovery, we used the receiver operating characteristic (ROC) curve to plot the false positive rate (FPR) against the true positive rate (TPR) defined, respectively, as,

$$\text{FPR} = \frac{\sum_{1\le j<k\le p} \boldsymbol{I}(\Omega_{0,jk} = 0, \hat{\Omega}_{jk} \neq 0)}{\sum_{1\le j<k\le p} \boldsymbol{I}(\Omega_{0,jk} = 0)}, \quad \text{TPR} = \frac{\sum_{1\le j<k\le p} \boldsymbol{I}(\Omega_{0,jk} \neq 0, \hat{\Omega}_{jk} \neq 0)}{\sum_{1\le j<k\le p} \boldsymbol{I}(\Omega_{0,jk} \neq 0)},$$

where $\hat{\Omega}$ denotes the estimated network. The $F1$ score [42], which is between 0 and 1, measures the accuracy of an estimator by summarizing both false positives and false negatives. Larger $F1$ scores indicate better structural recovery. For $\hat{\Omega}_{\lambda^*}$ estimated at the optimal penalty parameter λ^* selected by maximizing the $F1$ score, we also compared the entropy loss (EL) and Frobenius norm loss (FL) for estimation accuracy:

$$\text{EL} = \text{tr}(\Sigma_0 \hat{\Omega}_{\lambda^*}) - \log\det(\Sigma_0 \hat{\Omega}_{\lambda^*}) - p, \quad \text{FL} = \frac{\sum_{1\le j<k\le p}(\Omega_{0,jk} - \hat{\Omega}_{jk,\lambda^*})^2}{\sum_{1\le j<k\le p}(\Omega_{0,jk})^2}.$$

Our first comparison is based on only microbiome data where $p = q = 60$ and $n = 100$. In this example, the percentage of zeros per species ranges from 0% to 70%. Input for gCoda is the censored abundance matrix $\mathcal{D} = (\boldsymbol{r}^{(1)} + 0.5, \ldots, \boldsymbol{r}^{(n)} + 0.5)^\intercal$. The clr transformation is then applied to each row in $\mathcal{D}$ and the resulting matrix is used as input for SPIEC-EASI. Figure 4 shows the ROC curves obtained from different methods across different network models. One can see that SPIEC-EASI and gCoda perform similarly, and both underperform compared to metaMint. Because the nearest-neighbor network is denser, the ROC curves in the right panel of Fig. 4 are generally lower compared to their counterparts in other network models.

In our second study, we look at larger datasets where the number of metabolites is $q = 100$ and the number of microbes is $p - q = 100$. The sample size is $n = 300$. The method gCoda is thus not applicable because it was proposed specifically for microbiome data. Because we only censor microbiome data, the proportion of censored variables in this example is smaller. We first compare different methods in terms of network structural recovery. Figure 5 shows the average $F1$ score of each method across a range of penalty parameters. It can be seen that metaMint has overall higher F_1 scores than SPIEC-EASI, and closely resembles the oracle estimator.

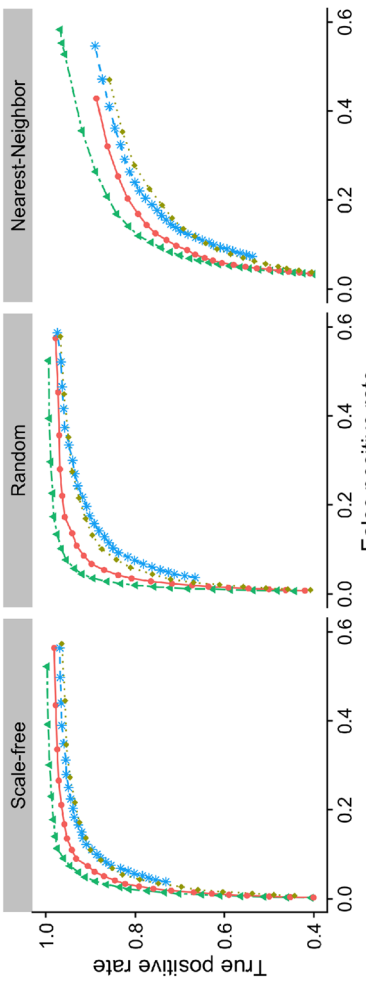

Fig. 4 ROC curves for the first study with $p = q = 60$ and $n = 100$: Oracle (two-dash line in green), metaMint (solid line in red), SPIEC-EASI (dotted line in brown), and gCoda (dashed line in blue). These results are averaged over 20 replications. metaMint outperforms SPIEC-EASI and gCoda

Since we know the true network structure, we also look at comparisons in terms of inverse covariance estimation accuracy at the optimal penalty parameter selected by maximizing the $F1$ score. As shown in Fig. 6, SPIEC-EASI performs the worst in all cases because its entropy and Frobenius norm loss are the largest. It is worth pointing out that there still exists substantial gap in both EL and FL between metaMint and the oracle estimator as a result of censoring. We anticipate that this issue can be partly addressed with increased sequencing depths.

4 Analysis of Bacterial Vaginosis Data

4.1 Data Description and Processing

Bacterial vaginosis (BV) is a common vaginal condition characterized by depletion of specific Lactobacillus species and increased abundance of diverse anaerobic bacteria such as genus *Gardnerella, Prevotella and others* [43, 44]. This condition affects an estimated 30% of women at any given time [45], and is associated with increased transmission of HIV and increased risk of preterm labor [46, 47]. Improved diagnosis and treatment of BV require not only a clearer understanding of the roles of BV associated bacterial species and their interactions, but also a detailed catalog of the interactions between these bacteria and relevant metabolites. We applied the proposed multi-omic approach to a cohort of 131 Rwandan women from McMillan et al. [9]. The microbiome data from sequencing the 16S rRNA gene consist of 51 bacterial species after initial filtering, and the vaginal metabolome determined by GC-MS contains 128 metabolites [see the Methods section in 9]. One bacterial species is present in only 13 individuals, so we removed this rare species and used 50 taxa in all analysis. Of the 131 women, 79 were normal, 23 were diagnosed with BV, 22 as being intermediate between BV and the normal state, and 7 did not have diagnosis. To account for the different sequencing depths, we applied the clr and modified clr to the microbiome data. Metabolomic data available from McMillan et al. [9] have already been log transformed. After the mclr transformation, a species is treated as censored at zero if it has at least one zero count. Based on this criterion, 27 of the 50 species are left censored.

We compare metaMint with SPIEC-EASI by applying the former to mclr transformed data and the latter to clr transformed data. At the optimal tuning parameter, which was selected using the stability approach in Liu et al. [35] with pre-specified stability threshold α, we randomly subsampled 80% of all samples to estimate the network using each method. This procedure was repeated 50 times and an edge selection frequency matrix was constructed such that each entry represents the proportion of times the corresponding edge was present. Only edges with at least 95% selection frequency were kept.

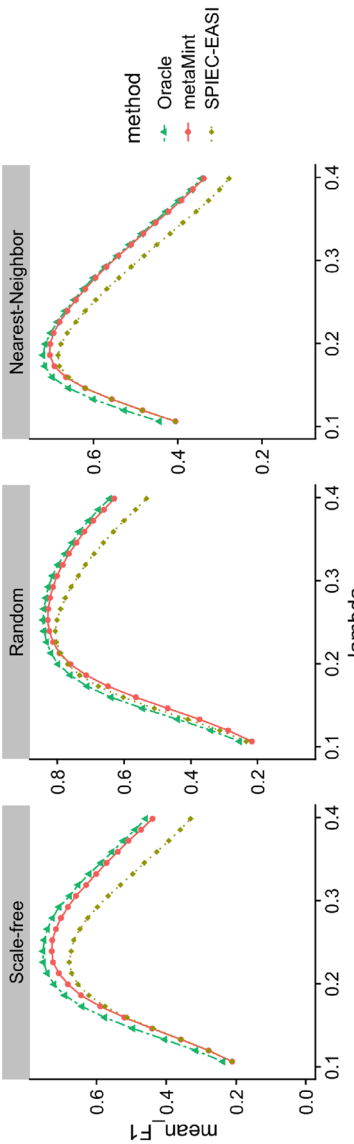

Fig. 5 Average $F1$ scores for different methods across different network models over 50 replications in the second simulation study: Oracle (two-dash line in green), metaMint (solid line in red), and SPIEC-EASI (dotted line in brown). gCoda is not applicable in this case because it was specific for microbiome data. metaMint outperforms SPIEC-EASI

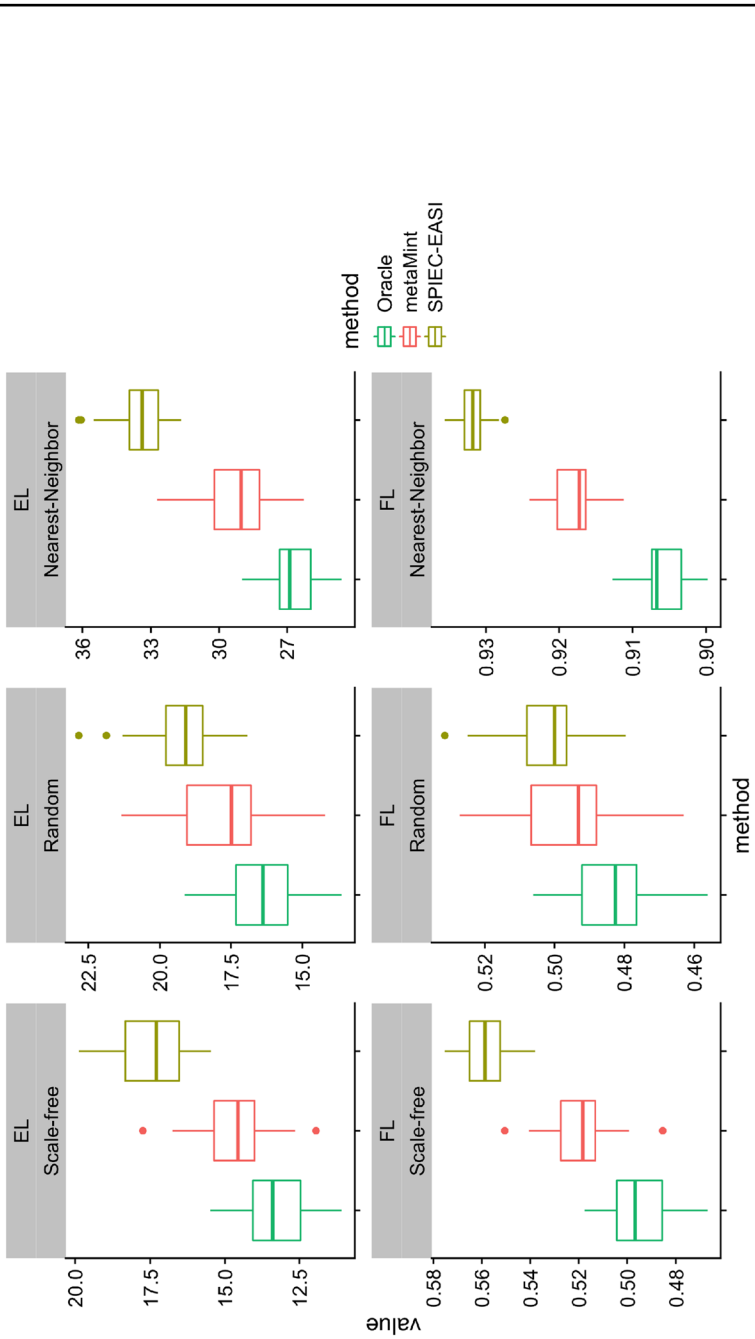

Fig. 6 Boxplots showing the entropy loss (EL, top row) and Frobenius norm loss (FL, bottom row) for different methods across different network models over 50 replications in the second simulation study. For each method, the optimal penalty parameter was selected as the one that maximizes $F1$ score. The oracle estimator performs the best, followed by metaMint, and SPIEC-EASI has the largest entropy and Frobenius norm loss

4.2 Results

We first compare metaMint and SPIEC-EASI by estimating a single integrated microbe and metabolite network for all subjects at stability threshold $\alpha = 0.01$. Figure 7 presents the joint microbe-metabolite network estimated by the two methods, where the thick black edges are shared between the two methods, blue edges are unique to metaMint, and red edges are unique to SPIEC-EASI. We can see that a majority of edges are shared between the two methods. In particular, both methods reported the conditional association between the genus *Gardnerella* and metabolite *GHB* (6–82), and between *Lactobacillus* and *unknown sugar 1* (3–165). These two edges are relatively stable and show up in the network for any stability threshold $\alpha \geq 0.004$. Importantly, the interaction between *Gardnerella* and *GHB* was also observed and reproduced experimentally in McMillan et al. [9]. Other notable microbe-metabolite interactions that are unique to each method include *Prevotella—unknown sugar 2* (7–166) estimated only by metaMint, and *Dialister—n-acetylputrescine* (10–106), *Dialister—phenylethylamine* (10–111) estimated only by SPIEC-EASI. These microbe-metabolite interactions are unique to each method until the stability threshold increases to $\alpha = 0.02$. The differences reported by the two methods are manifestations of the different transformations and whether the model directly accounts for zero inflation.

To gain further insights into the roles of these microbe-metabolite interactions, we partitioned all subjects into two groups: the normal group ($n_1 = 79$) and everyone else (the BV group, $n_2 = 52$). metaMint and SPIEC-EASI were applied to estimate a network for each group using the same model selection procedure as before. In general, we observe more interactions in the group-specific network estimated by SPIEC-EASI compared to the corresponding network estimated by metaMint. At stability threshold 0.01, no interaction between microbes and metabolites was recovered due to the reduced sample size in each group. As we gradually increase the stability threshold, the first pair of microbe-metabolite interaction unique to the BV group is between *Gardnerella* and *GHB*, and was identified by both metaMint and SPIEC-EASI. Table 1 provides a list of microbe-metabolite interactions that are unique to each group of patients identified by both methods at stability threshold 0.02. It is worth noting that *Gardnerella—GHB*, *Prevotella—unknown sugar 2*, and *Dialister—cadaverine* only show up for the BV group, whereas the interactions between *Lactobacillus* species and several metabolites appear only for the normal group. Abundance of *Lactobacillus* and *Prevotella* has long been used as a diagnostic signature for bacterial vaginosis [43, 44]. In addition, McMillan et al. [9] hypothesized that *Dialister* is responsible for malodor in the vagina. Our analysis may shed light on the mechanistic link between metabolic end products and microbes in vaginal bacterial communities, and provide key guidance regarding the diagnosis and treatment of BV.

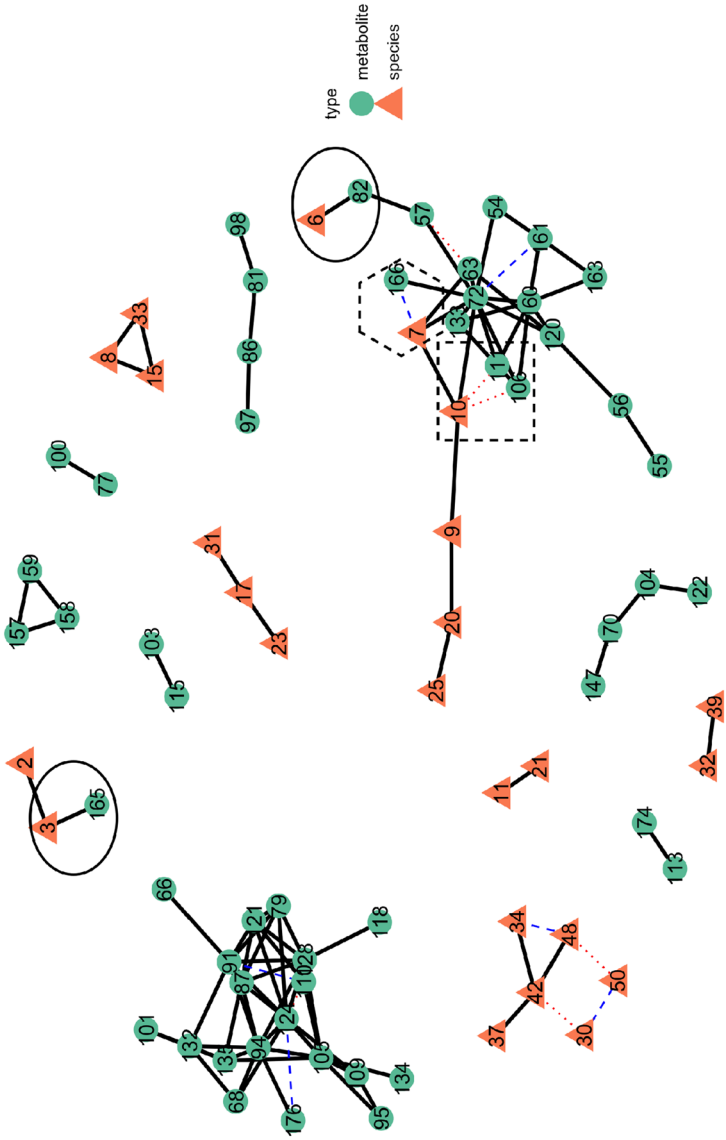

Fig. 7 The overlayed microbe-metabolite network in the BV data example estimated from metaMint and SPIEC-EASI. Color and shape of each node indicate whether the node is a metabolite or bacterial species. Thick black edges are shared between the two methods, whereas dashed blue edges are unique to metaMint and dotted red edges are unique to the SPIEC-EASI

Table 1 Microbe-metabolite interactions estimated by metaMint and SPIEC-EASI that are unique to each group

Microbe id	Metabolite id	Microbe name	Metabolite name	Group
1	62	Lactobacillus_iners	2-O-Glycerol-D-galactopyranoside 3	Normal
2	88	Lactobacillus_crispatus	Glyceric_acid	Normal
2	125	Lactobacillus_crispatus	Succinate	Normal
3	96	Lactobacillus	Malate	Normal
3	125	Lactobacillus	Succinate	Normal
6	82	Gardnerella	GHB	BV
7	166	Prevotella	Unknown sugar 2	BV
10	72	Dialister	Cadaverine	BV

5 Discussion

The uneven sequencing depths and sparsity in microbiome data present significant challenges in inferring interactions between microbial species and their products. The different sequencing depths imply different levels of uncertainty, but how to handle varying sequencing depths in multivariate statistical analysis remains an unsolved problem [48, 49]. This paper proposes the censored Gaussian graphical model for joint estimation of microbiome and metabolomic network, which can be used to identify conditional dependencies (direct interactions) between microbial species and metabolites. Key to our proposal is the use of the modified centered log-ratio for transforming the observed microbial counts, which is scale invariant and preserves the ranking of positive counts relative to zeros. Observed zeros are attributed to undersampling and modeled as due to left censoring. Our method metaMint can be generalized to study other omics data types that fit in the censored Gaussian graphical model framework. Analysis of the bacterial vaginosis data demonstrates that metaMint facilitates the discovery of important microbe-metabolite interactions for diagnosis and treatment of this condition. The data example in Sect. 4 has about 50% censored variables, although 11 of them have less than 10% zero counts. As we move into high-resolution studies which collect microbiome data at the strain or amplicon sequence variant level, our model that explicitly accounts for observed zeros may exhibit more advantage over existing methods.

From a methodological perspective, metaMint estimates the correlations in a marginal manner, which may not be optimal because marginal approaches ignore the fact that the correlation matrix is positive semi-definite. Augugliaro et al. [26] proposed an approximated EM algorithm that jointly estimates all entries in the correlation matrix; however, their method only works well under specific settings and there is a lack of theoretical understanding about the resulting estimator. Obvious but non-trivial extension is to explore computationally and statistically efficient alternatives that jointly estimate all entries in the correlation matrix.

Our model is related to but substantially different from the zero-inflated Gaussian graphical model in McDavid et al. [27]. While our model assumes the observed

zeros are due to undersampling, McDavid et al. [27] uses a two-part Hurdle model that treats all zeros as structural. The multivariate Hurdle model consists of an Ising model that captures the discrete part and a Gaussian graphical model that describes the continuous part if the hurdle is passed. When the study design favors the two-part process, as is the case in single-cell RNA-seq analysis, the multivariate Hurdle model should be considered. On the other hand, the censored Gaussian graphical model is simpler and works well if the study design favors sampling zeros and/or structural zeros can be reasonably approximated as sampling zeros [21].

It is worth pointing out that the observed data defined in (1) are continuous-valued. In this paper, we have made the simplifying assumption that the observed counts can be approximated by a log-normal distribution with left censoring. An alternative approach is to analyze observed counts directly while still treating zeros as due to left censoring. In the regression setting, Clark et al. [50] provided a general framework that uses a latent continuous variable to model observed species abundance, which can be presence/absence, continuous abundance, ordinal counts, or counts that are subject to a total sum constraint. It would be interesting to see if similar ideas can be used to model interactions between microbial species and other molecules.

Acknowledgements J. Ma is partially supported by NIH 1R01GM129512-01. The author would like to thank three anonymous referees for their constructive comments and suggestions.

References

1. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, FitzGerald MG, Fulton RS et al (2012) Structure, function and diversity of the healthy human microbiome. Nature 486(7402):207–214
2. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, Brady A, Creasy HH, McCracken C, Giglio MG et al (2017) Strains, functions and dynamics in the expanded human microbiome project. Nature 550(7674):61–66
3. Callahan BJ, McMurdie PJ, Holmes SP (2017) Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J 11(12):2639–2643
4. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R (2018) Current understanding of the human microbiome. Nat Med 24(4):392
5. iHMP Research Network Consortium (2019) The integrative human microbiome project. Nature 569:641–648
6. McHardy IH, Goudarzi M, Tong M, Ruegger PM, Schwager E, Weger JR, Graeber TG, Sonnenburg JL, Horvath S, Huttenhower C et al (2013) Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. Microbiome 1(1):17

7. Wu GD, Compher C, Chen EZ, Smith SA, Shah RD, Bittinger K, Chehoud C, Albenberg LG, Nessel L, Gilroy E et al (2016) Comparative metabolomics in vegans and omnivores reveal constraints on diet-dependent gut microbiota metabolite production. Gut 65(1):63–72
8. Jia W, Xie G, Jia W (2018) Bile acid-microbiota crosstalk in gastrointestinal inflammation and carcinogenesis. Nat Rev Gastroenterol Hepatol 15(2):111–128
9. McMillan A, Rulisa S, Sumarah M, Macklaim JM, Renaud J, Bisanz JE, Gloor GB, Reid G (2015) A multi-platform metabolomics approach identifies highly specific biomarkers of bacterial diversity in the vagina of pregnant and non-pregnant women. Sci Rep 5:14174
10. Org E, Blum Y, Kasela S, Mehrabian M, Kuusisto J, Kangas AJ, Soininen P, Wang Z, Ala-Korpela M, Hazen SL et al (2017) Relationships between gut microbiota, plasma metabolites, and metabolic syndrome traits in the metsim cohort. Genome Biol 18(1):70
11. Liu R, Hong J, Xu X, Feng Q, Zhang D, Gu Y, Shi J, Zhao S, Liu W, Wang X et al (2017) Gut microbiome and serum metabolome alterations in obesity and after weight-loss intervention. Nat Med 23(7):859–868
12. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, Andrews E, Ajami NJ, Bonham KS, Brislawn CJ et al (2019) Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature 569(7758):655–662
13. Gould AL, Zhang V, Lamberti L, Jones EW, Obadia B, Gavryushkin A, Korasidis N, Carlson JM, Beerenwinkel N, Ludington WB (2018) High-dimensional microbiome interactions shape host fitness. Proc Natl Acad Sci 115(51):E11951–E11960
14. Friedman J, Alm EJ (2012) Inferring correlation networks from genomic survey data. PLoS Comput Biol 8(9):e1002687
15. Fang H, Huang C, Zhao H, Deng M (2015) CCLasso: correlation inference for compositional data through lasso. Bioinformatics 31(19):3172–3180
16. de la Fuente A, Bing N, Hoeschele I, Mendes P (2004) Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics 20(18):3565–3574
17. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA (2015) Sparse and compositionally robust inference of microbial ecological networks. PLoS Comput Biol 11(5):e1004226
18. Friedman JH, Hastie TJ, Tibshirani RJ (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9(3):432–441
19. Fang H, Huang C, Zhao H, Deng M (2017) gCoda: conditional dependence network inference for compositional data. J Comput Biol 24(7):699–708
20. Kaul A, Mandal S, Davidov O, Peddada SD (2017) Analysis of microbiome data in the presence of excess zeros. Front Microbiol 8:2114
21. Silverman JD, Roche K, Mukherjee S, David LA (2018) Naught all zeros in sequence count data are the same. bioRxiv, p 477794
22. Yoon G, Gaynanova I, Müller CL (2019) Microbial networks in SPRING-semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data. Front Genet 10:516
23. Hoffman HJ, Johnson RE (2015) Pseudo-likelihood estimation of multivariate normal parameters in the presence of left-censored data. J Agric Biol Environ Stat 20(1):156–171
24. Pesonen M, Pesonen H, Nevalainen J (2015) Covariance matrix estimation for left-censored data. Comput Stat Data Anal 92:13–25
25. Jones MP, Perry SS, Thorne PS (2015) Maximum pairwise pseudo-likelihood estimation of the covariance matrix from left-censored data. J Agric Biol Environ Stat 20(1):83–99
26. Augugliaro L, Abbruzzo A, Vinciotti V (2018) ℓ_1-penalized censored gaussian graphical model. Biostatistics 21:1–16
27. McDavid A, Gottardo R, Simon N, Drton M et al (2019) Graphical models for zero-inflated single cell gene expression. Ann Appl Stat 13(2):848–873
28. Guo J, Levina E, Michailidis G, Zhu J (2015) Graphical models for ordinal data. J Comput Gr Stat 24(1):183–204
29. Suggala AS, Yang E, Ravikumar P (2017) Ordinal graphical models: a tale of two approaches. In: International conference on machine learning, pp 3260–3269
30. Tobin J (1958) Estimation of relationships for limited dependent variables. Econom: J Econom Soc 26(1):24–36
31. Henningsen A (2010) Estimating censored regression models in R using the `censreg` package. R package vignettes

32. Fan J, Liu H, Ning Y, Zou H (2017) High dimensional semiparametric latent graphical model for mixed data. J R Stat Soc: Ser B (Stat Methodol) 79(2):405–421
33. Cai TT, Liu W, Luo X (2011) A constrained ℓ_1 minimization approach to sparse precision matrix estimation. J Am Stat Assoc 106(494):594–607
34. Cai TT, Liu W, Zhou HH (2016) Estimating sparse precision matrix: optimal rates of convergence and adaptive estimation. Ann Stat 44(2):455–488
35. Liu H, Roeder K, Wasserman L (2010) Stability approach to regularization selection (stars) for high dimensional graphical models. In: Advances in neural information processing systems, pp 1432–1440
36. van den Boogaart KG, Tolosana-Delgado R (2013) Analyzing compositional data with R, vol 122. Springer, Berlin
37. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ (2017) Microbiome datasets are compositional: and this is not optional. Front Microbiol 8:2224
38. Zhou W, Sailani MR, Contrepois K, Zhou Y, Ahadi S, Leopold SR, Zhang MJ, Rao V, Avina M, Mishra T et al (2019) Longitudinal multi-omics of host-microbe dynamics in prediabetes. Nature 569(7758):663–671
39. Vandeputte D, Kathagen G, D'hoe K, Vieira-Silva S, Valles-Colomer M, Sabino J, Wang J, Tito RY, De Commer L, Darzi Y et al (2017) Quantitative microbiome profiling links gut community variation to microbial load. Nature 551(7681):507–511
40. Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512
41. Erdős P, Rényi A (1960) On the evolution of random graphs. Publ Math Inst Hung Acad Sci 5:17–61
42. van Rijsbergen CJ (1979) Information retrieval, 2nd edn. Butterworth-Heinemann, Newton
43. Fredricks DN, Fiedler TL, Marrazzo JM (2005) Molecular identification of bacteria associated with bacterial vaginosis. N Engl J Med 353(18):1899–1911
44. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SS, McCulle SL, Karlebach S, Gorle R, Russell J, Tacket CO et al (2011) Vaginal microbiome of reproductive-age women. Proc Natl Acad Sci 108(Supplement 1):4680–4687
45. Koumans EH, Sternberg M, Bruce C, McQuillan G, Kendrick J, Sutton M, Markowitz LE (2007) The prevalence of bacterial vaginosis in the united states, 2001–2004; associations with symptoms, sexual behaviors, and reproductive health. Sex Transm Dis 34(11):864–869
46. Guerra B, Ghi T, Quarta S, Morselli-Labate AM, Lazzarotto T, Pilu G, Rizzo N (2006) Pregnancy outcome after early detection of bacterial vaginosis. Eur J Obstet Gynecol Reprod Biol 128(1–2):40–45
47. Atashili J, Poole C, Ndumbe PM, Adimora AA, Smith JS (2008) Bacterial vaginosis and hiv acquisition: a meta-analysis of published studies. AIDS 22(12):1493
48. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR, Vázquez-Baeza Y, Birmingham A et al (2017) Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome 5(1):27
49. McKnight DT, Huerlimann R, Bower DS, Schwarzkopf L, Alford RA, Zenger KR (2019) Methods for normalizing microbiome data: an ecological perspective. Methods Ecol Evol 10(3):389–400
50. Clark JS, Nemergut D, Seyednasrollah B, Turner PJ, Zhang S (2017) Generalized joint attribute modeling for biodiversity analysis: median-zero, multivariate, multifarious data. Ecol Monogr 87(1):34–56