frontiers | Frontiers in Microbiomes

TYPE Original Research
PUBLISHED 03 June 2024
DOI 10.3389/frmbi.2024.1366948

OPEN ACCESS

EDITED BY
Muhammad Ali,
Trinity College Dublin, Ireland

REVIEWED BY
Nar Singh Chauhan,
Maharshi Dayanand University, India
Yuan Jiang,
Oregon State University, United States

*CORRESPONDENCE
Anne G. Hoen
anne.g.hoen@dartmouth.edu

RECEIVED 07 January 2024
ACCEPTED 30 April 2024
PUBLISHED 03 June 2024

CITATION
Zhou J, Gui J, Viles WD, Chen H, Li S, Madan JC, Coker MO and Hoen AG (2024) Identifying stationary microbial interaction networks based on irregularly spaced longitudinal 16S rRNA gene sequencing data. *Front. Microbiomes* 3:1366948. doi: 10.3389/frmbi.2024.1366948

Identifying stationary microbial interaction networks based on irregularly spaced longitudinal 16S rRNA gene sequencing data

Jie Zhou[1], Jiang Gui[1], Weston D. Viles[2], Haobin Chen[1], Siting Li[1], Juliette C. Madan[3], Modupe O. Coker[3,4] and Anne G. Hoen[1,3]*

[1]Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Hanover, NH, United States, [2]Khoury College of Computer Science, Northeastern University, Portland, ME, United States, [3]Department of Epidemiology, Geisel School of Medicine, Dartmouth College, Hanover, NH, United States, [4]School of Dental Medicine, University of Rutgers, Newark, NJ, United States

Introduction: The microbial interactions within the human microbiome are complex, and few methods are available to identify these interactions within a longitudinal microbial abundance framework. Existing methods typically impose restrictive constraints, such as requiring long sequences and equal spacing, on the data format which in many cases are violated.

Methods: To identify microbial interaction networks (MINs) with general longitudinal data settings, we propose a stationary Gaussian graphical model (SGGM) based on 16S rRNA gene sequencing data. In the SGGM, data can be arbitrarily spaced, and there are no restrictions on the length of data sequences from a single subject. Based on the SGGM, EM -type algorithms are devised to compute the $L1$-penalized maximum likelihood estimate of MINs. The algorithms employ the classical graphical LASSO algorithm as the building block and can be implemented efficiently.

Results: Extensive simulation studies show that the proposed algorithms can significantly outperform the conventional algorithms if the correlations among the longitudinal data are reasonably high. When the assumptions in the SGGM areviolated, e.g., zero inflation or data from heterogeneous microbial communities, the proposed algorithms still demonstrate robustness and perform better than the other existing algorithms. The algorithms are applied to a 16S rRNA gene sequencing data set from patients with cystic fibrosis. The results demonstrate strong evidence of an association between the MINs and the phylogenetic tree, indicating that the genetically related taxa tend to have more/stronger interactions. These results strengthen the existing findings in literature.

Discussion: The proposed algorithms can potentially be used to explore the network structure in genome, metabolome etc. as well.

KEYWORDS

Gaussian graphical model, microbial interaction network, EM algorithm, longitudinal data, relative abundance

1 Introduction

Microorganisms thrive in communities in large numbers. They interact with their host and with one another in various ways, such as commensalism, synergism, competition, parasitism, and predation. This complex set of interactions can be depicted in the form of microbial interaction networks (MINs) (Faust and Raes, 2012). Traditionally, such interactions have been inferred using culture-based methods, which can only accommodate a small number of microbial strains (Gause, 1934; Staley and Konopka, 1985; Harcombe, 2010). Since most microbes cannot be cultivated, the estimated interactions under laboratory conditions could be misleading. Underpinned by advances in next-generation sequencing (NGS) technologies, a complete microbiome profile can be measured at a relatively low cost, allowing researchers to investigate microbial interactions in situ. However, the complexities of these high-throughput data, such as their high dimensionality, zero inflation, and compositional nature, pose substantial challenges to identifying MINs (Faust and Raes, 2012). Currently, the primary way to infer MINs is the pairwise method, in which the cooccurrence or mutual exclusion pattern of two species is compared using measures such as Pearson or Spearman correlation (Qin et al., 2010; Zhou et al., 2010; Arumugam et al., 2011; Barberan et al., 2012). An emerging method is based on conditional independence, i.e., the conditional joint distribution of two taxa given all the other microbiome members. Conditional independence is conceptually superior to the pairwise method since it removes the effects of all the other taxa when measuring the relationship between the two taxa of interest (Kurtz et al., 2015; Chen et al., 2017; Viles et al., 2021). Furthermore, if the data follow a normal distribution, then the precision matrix, i.e., the inverse of the covariance matrix, directly reflects the conditional independence relationship among microbes. With such an appealing interpretation, precision matrices have become the ideal tools for exploring the structure of MINs (Fang et al., 2017; Yoon et al., 2019; Yuan et al., 2019; Jiang et al., 2020; Tian et al., 2023).

In particular, the authors in (Kurtz et al., 2015) proposed a conditional independence-based pipeline named SParse InversE Covariance Estimation for Ecological Association Inference (SPIEC-EASI) to estimate MINs. In SPIEC-EASI, $L1$-penalized maximum likelihood estimation of the precision matrix is employed to identify high-dimensional MINs. Mathematically, the $L1$-penalized maximum likelihood estimation of the precision matrix has been studied extensively in the literature (Yuan and Lin, 2007; Avella-Medina et al., 2018; Wang and Jiang, 2020). Algorithms have been proposed to compute such estimates, e.g., graphical LASSO (Friedman et al., 2008; Friedman et al., 2019) and the neighborhood method (Meinshansen and Bühlmann, 2006). In SPIEC-EASI, graphical LASSO computes the precision matrix recursively based on the coordinate descent algorithm. In contrast, the neighborhood method computes the neighborhood of each node and then combines these neighborhoods to form an estimate of the network. However, a prerequisite of SPIEC-EASI is that the data should be independent. Although independence is a reasonable assumption if the data are cross-sectional, in many other cases, data sets are longitudinal, in which multiple observations are made on the same subject. In such studies, the observations from the same subject are typically correlated and violate the assumption of SPIEC-EASI. There have been studies to estimate the network from the correlated data. The time series models, e.g., vector autoregression, have been employed to address the correlation between observations within the same cluster (Bach and Jordan, 2004; Qiu et al., 2016; Chen et al., 2017; Epskamp et al., 2018; He et al., 2022). In particular, the authors in (He et al., 2022) used autoregression in the proposed ARZIMM model to characterize the longitudinal absolute abundance data for the microbiome study. However, time series methods require the data for each subject to be long enough and equally spaced, which is not usually satisfied in reality. Functional data analysis has also been used to decipher the conditional correlation for high-dimensional data (Zhu et al., 2016; Li and Solea, 2018; Qiao et al., 2019; Solea and Li, 2020). For example, for electroencephalogram (EEG) data (Qiao et al., 2019), proposed a functional graphical model to estimate a network of brain reactions, and (Solea and Li, 2020) proposed the copula Gaussian graphical model for a network of functional magnetic resonance imaging (fMRI) data. Functional data analysis based methods require the data to be densely spaced and the sample size to be large. The requirements inherited in these existing methods are often violated for longitudinal data sets in human microbiome studies.

In this paper, we consider the estimation of MINs from irregularly spaced longitudinal 16S rRNA gene sequencing data. The SPIEC-EASI pipeline can be seen as a special case of the proposed algorithms. The inferences are considered under three conditions. In the first condition, we assume that all subjects share an autocorrelation parameter τ during the trial. For this case, we propose a model named the homogeneous SGGM to characterize MINs. For the homogeneous SGGM, a recursive graphical LASSO algorithm is proposed to compute the $L1$-penalized maximum likelihood estimate (MLE) of the network. In the second condition, the homogeneous SGGM is extended to the heterogeneous SGGM, allowing different subjects to have their own autocorrelation parameter. For the heterogeneous SGGM, an expectation-maximization (EM)-type algorithm is devised to compute the $L1$-penalized MLE of the network. In the third condition, the autocorrelation parameters are further allowed to depend on covariates such as sex and race. We show how the algorithm in condition two can be adapted to accommodate the extension. Extensive simulation studies are conducted to compare the proposed algorithms with existing algorithms, including the SPIEC-EASI pipeline and the GGMselect algorithm family (Giraud et al., 2012). The comparisons are conducted under different scenarios, aiming to investigate the robustness of the algorithms to violations of the assumptions of the SGGM. This is necessary since the 16S rRNA gene sequencing data are highly irregular and may fail to exactly satisfy the premises of the proposed models and algorithms. For all the scenarios considered, the proposed algorithms exhibit better performance for network selection than that of other existing algorithms.

In the final part, the proposed models are employed to study a longitudinal gut microbiome data set from a cohort with cystic fibrosis in New Hampshire (Madan et al., 2012). To validate the proposed algorithms, with the estimated MINs, we measure the

correlation between the estimated MINs and the corresponding phylogenetic tree. A permutation test is proposed to determine the significance of such a correlation. The results demonstrate strong evidence for the positive correlation between the MINs and the phylogenetic tree, indicating that genetically related taxa also tend to have more/stronger interactions. These findings strengthen the discoveries that have been reported in other studies (Chaffron et al., 2010; Eiler et al., 2012) and provide an empirical basis for using phylogenetic trees as a tool to explore microbial interactions in future studies (Chung et al., 2022).

The paper is organized as follows. In the *Materials and methods* section, we introduce stationary Gaussian graphical models (SGGMs) and three related inference algorithms. In the *Results* section, we compare the performance of the proposed algorithms with that of the conventional methods under different scenarios and demonstrate the superiority of the proposed algorithm. We then considered the gut microbiome of subjects with cystic fibrosis. The homogeneous version of the proposed algorithm is employed to identify the MINs of the microbiome. The plausibility of the estimated MINs is discussed. The *Discussion* section includes a brief review of the models.

2 Materials and methods

2.1 Data generation process

Let $y_{it_{ik}} = (y_{it_{ik}1}, \cdots, y_{it_{ik}p})^T$ denote observations of some transformed abundance data of a microbiome with p taxa from subject i at time t_k $(1 \le i \le m, 1 \le k \le n_i)$ so that it is appropriate to assume that $y_{it_{ik}} \sim N_p(\mu, \Sigma)$, where $\mu = E(y_{it_{ik}})$ and $\Sigma = Var(y_{it_{ik}})$. The precision matrix is defined as $\Omega = \Sigma^{-1}$. Then, the $n_i p$ vector $\mathrm{y}_i = (\mathrm{y}_{i1}^T, \cdots, \mathrm{y}_{in_i}^T)^T$ represents all the observations on subject i, and vector $\mathbf{y} = (\mathrm{y}_1^T, \cdots, \mathrm{y}_m^T)^T$ represents the observations on all the m subjects with $n = \sum_{i=1}^m n_i$. For the correlations between the observations, we assume that the observations from different subjects are independent, i.e., $cov(y_{i_1 t_{ik_1}}, y_{i_2 t_{ik_2}}) = 0_{p\times p}$ for $i_1 \ne i_2, k_1 \ge 1, k_2 \ge 1$. For observations from the 1 same subject, we assume $cov(y_{it_{ik_1}}, y_{it_{ik_2}}) = DH_{ik_1k_2}D$ where $D = \mathrm{diag}(\sigma_1, \cdots, \sigma_p)$ with $\sigma_1^2, \cdots, \sigma_p^2$ diagonal elements of Σ, while $H_{ik_1}k_2$ is the correlation matrix between $Y_{it_{ik_1}}$ and $Y_{it_{ik_2}}$ for which the following form is assumed:

$$H_{ik_1k_2} = \Phi_{ik_1k_2} \odot R \tag{1}$$

The symbol in (Equation 1) stands for the Hadamard product of matrices $\Phi_{ik_1k_2}$ and R. Here, R is the correlation matrix with respect to covariance matrix Σ, while matrix $\Phi_{ik_1k_2} = (\phi_{ik_1k_2})_{p\times p}$ defines the dampening rates at which the components of $H_{ik_1k_2}$ decrease as time goes from t_{ik_1} to t_{ik_2}. For example, $(\Phi_{ik_1k_2})_{12}$ is the dampening rate of correlation $cor(Y_{it_{ik_1}1}, Y_{it_{ik_1}2})$ to correlation $cor(Y_{it_{ik_1}1}, Y_{it_{ik_2}2})$. Theoretically, dampening rates can vary from taxon to taxon and depend on the time points as long as the resulting matrix $H_{ik1}k_2$ is positive definite. However, in this paper, for subject i, we assume that the components of $H_{ik_1k_2}$ have the same dampening rate. Furthermore, they depend on time points (t_{ik_1}, t_{ik_2}) only through the distance between t_{ik_1} to t_{ik_2}, i.e., $\phi_{ik_1k_2} = g_i(|t_{ik_1} - t_{ik_2}|)$ for some decreasing function $0 \le g_i(\cdot) \le 1$. Motivated by studies on the longitudinal regression model (Diggle et al., 2002), we assume function $g_i(\cdot)$ has the form of $\exp(-\tau_i|t_{ik_1} - t_{ik_2}|^p)$. For $p = 0$, we have $\phi_{ik_1k_2} = \exp(-\tau_i)$, which is referred to as the uniform correlation and can be used to model the spatial correlation. For example, specimens may be collected at different body sites from the same subjects, for which the uniform correlation seems to be a reasonable assumption. On the other hand, the cases of $p > 0$ can be used to model the irregularly spaced temporal correlation, which typically decreases as the time span $|t_{ik_1} - t_{ik_2}|$ increases. In particular, functions $\exp(\tau_i|t_{ik_1} - t_{ik_2}|)$ and $\exp(\tau_i|t_{ik_1} - t_{ik_2}|^2)$ have been used in the marginal regression model for low-dimensional longitudinal data. Here, the parameters τ_is, which are referred to as autocorrelation parameters, measure the dampening rates that are shared by all the components of y_{it}, $(i = 1, \cdots, m)$.

Without loss of generality, we always employ the correlation function $\exp(\tau_i|t_{ik_1} - t_{ik_2}|)$ in the following and assume that the observations have been centered so that $\mu = 0$. Let Σ_i denote the covariance matrix of the observation vector $\boldsymbol{y}_i$. The density function of y is then given by

$$f(y|\Omega, \tau) = \prod_{i=1}^m f_i(y_i|\Sigma_i, \tau),$$

$$\text{where} \quad f_i(y_i|\Sigma_i, \tau_i) = (2\pi)^{-n_i p/2}|\Sigma_i|^{-1/2}\exp(-y_i^T\Sigma_i^{-1}y_i/2) \text{with}$$

$$\Sigma_i = \begin{pmatrix} \Omega^{-1} & e^{-\tau_i|t_{i1}-t_{i2}|}\Omega^{-1} & \cdots & e^{-\tau_i|t_{i1}-t_{in_1}|}\Omega^{-1} \\ e^{-\tau_i|t_{i2}-t_{i1}|}\Omega^{-1} & \Omega^{-1} & \cdots & e^{-\tau_i|t_{i2}-t_{in_2}|}\Omega^{-1} \\ \vdots & \vdots & \vdots & \vdots \\ e^{-\tau_i|t_{in_i}-t_{i1}|}\Omega^{-1} & e^{-\tau_i|t_{in_i}-t_{i2}|}\Omega^{-1} & \cdots & \Omega^{-1} \end{pmatrix} \tag{2}$$

Since the number of unknown parameters in Ω is much larger than the sample size in the context of the gut microbiome, the maximum likelihood estimate of Ω is unidentifiable, and sparsity is typically assumed in the literature. To this end, penalized maximum likelihood estimation (MLE) is usually adopted, e.g., the SPIEC-EASI model in (Kurtz et al., 2015). The SPIEC-EASI pipeline employes the L_1-penalty to achieve the sparsity of Ω for cross-sectional observations. Here, we adopt the same strategy for longitudinal data and use the minimizer of the following L_1-penalized negative log-likelihood function as the estimate of network Ω

$$(\hat{\Omega}, \hat{\tau}) = \arg\min_{\Omega,\tau}\{-2\log(f(y|\Omega, \tau)) + n\lambda|\Omega|_1\} \tag{3}$$

We refer to model (Equations 2, 3) the stationary Gaussian graphical model (SGGM). Here, stationarity stems from the fact that the same network Ω is shared by all the subjects and at all time points. If the data are independent observations, then (3) can be solved by the graphical LASSO algorithm (Friedman et al., 2008) or the neighborhood method (Meinshansen and Bühlmann, 2006), and SGGM is just reduced to the SPIEC-EASI model. However, since the data are longitudinal and can correlate to each other, the performance of the SPIEC-EASI pipeline is not guaranteed when solving (3).

Notably, in model (2, 3), we assume the same correlation dampening rate τ_i for all the taxa in the microbiome of subject i.

This assumption is motivated by the characteristics of the gut microbiome, where the taxa are typically influenced by the same perturbation sources, such as diet change and disease development. However, even if this assumption is violated and different taxa have different dampening rates, model (2, 3) can still be used as a working model for identifying the network structure, in which τ_i can be regarded as the mean dampening rate of the whole microbiome. In such cases, models (2, 3) still outperform the SPIEC-EASI pipeline, and the latter ignores the correlation structure of longitudinal data. We demonstrate this point through simulation studies in Section 3.1.

In the following sections, we propose three algorithms to identify the network Ω in (3) based on different dampening rate τ_i models. An algorithm for the homogeneous SGGM is first considered, and two extensions are proposed that allow the algorithms to deal with cohorts of heterogeneous subjects. These algorithms integrate the graphical LASSO algorithm with other algorithms, e.g., the EM algorithm, to find the penalized maximum likelihood estimator of Ω.

2.2 Homogeneous SGGM

In this section, we consider identifying Ω under the assumption $\tau_1 = \cdots = \tau_m = \tau$.

Thus, it is assumed that correlations between observations at different time points dampen at the same rate for each subject in the cohort. From the density function (2), the log-likelihood function for $y = (y_1^T, \cdots y_m^T)^T$ is given by

$$l_n(\Sigma, \tau|y) = -\frac{1}{2}\sum_{i=1}^{m}(p\log(|\Phi_i|) + n_i\log(|\Sigma|) + y_i^T(\Phi_i \otimes \Sigma)^{-1}y_i) \quad (4)$$

up to a constant. Here, $\otimes$ stands for the Kronecker product. We use the formulas $\Sigma_i = \Phi_i \otimes \Sigma$ and $|\Sigma_i| = |\Phi_i|^p|\Sigma|^{n_i}$. Note that with formula $(\Phi_i \otimes \Sigma)^{-1} = \Phi_i^{-1} \otimes \Sigma^{-1}$, the last term in (Equation 4) can be rewritten as

$$\begin{aligned} y_i^T(\Phi_i \otimes \Sigma)^{-1}y_i &= y_i^T(\Phi_i^{-1} \otimes \Sigma^{-1})y_i = \sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\phi_{ijk}^- y_{it_{ij}}^T \Omega\, y_{it_{ik}} \\ &= \sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\mathrm{tr}(\phi_{ijk}^- y_{it_{ik}} y_{it_{ij}}^T \Omega) = \mathrm{tr}\left(\left(\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\phi_{ijk}^- y_{it_{ik}} y_{it_{ij}}^T\right)\Omega\right) \\ &\triangleq \mathrm{tr}S_i(\tau)\,\Omega, \end{aligned} \quad (5)$$

where $S_i(\tau) = \sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\phi_{ijk}^- y_{it_k} y_{it_{ij}}^T$ and $\Phi_i^{-1} = (\phi_{ijk}^-)n_i \times n_i$. By substituting (Equation 5) into (4), we have

$$l_n(\Omega, \tau|y) = -\frac{1}{2}\left\{\sum_{i=1}^{m} p\log(|\Phi_i|) - n\log(|\Omega|) + n\,\mathrm{tr}(\bar{S}(\tau)\,\Omega)\right\} \quad (6)$$

where $n = \sum_{i=1}^{m} n_i, \bar{S}(\tau) = \frac{1}{n}\sum_{i=1}^{m} S_i(\tau)$. Here, we use $\bar{S}(\tau)$ to emphasize that matrix $\bar{S}$ is a function of unknown parameter τ. With (Equation 6) in hand, the sparse network can be achieved by minimization the L_1-penalized negative log-likelihood function (3), i.e.,

$$\min_{\Omega,\tau}\{-2l_n(\Omega, \tau|y) + n\lambda|\Omega|_1\} \quad (7)$$

for given tuning parameter $\lambda > 0$. The minimization problem (Equation 7) can be solved through a block coordinate descent procedure. First note that for a given τ, the solution of Ω can be obtained through the following minimization:

$$\min_{\Omega}\{-\log|\Omega| + tr(\bar{S}(\tau)\,\Omega) + \lambda|\Omega|_1\} \quad (8)$$

which has the same form as the GGM for independent data when the empirical covariance matrix is given by $\bar{S}(\tau)$. Consequently, the graphical LASSO algorithm can be used to compute the sparse estimate of Ω in (Equation 8). On the other hand, given Ω, the minimization of (7) with respect to τ does not involve any L_1 penalty term and consequently can be carried out through the maximization of the likelihood function (6) with respect to τ. The conventional Newton algorithm can be used in this step. This process continues until convergence is achieved. This algorithm will be referred to as homogeneous longitudinal graphical LASSO (LGLASSO), for which the details are summarized in the following table.

1: **procedure** Given Initial Value τ_0 And Ω_0, Tuning Parameter λ And Error Tolerance $e > 0$:

2: With $\tau = \tau_0$, solve optimization problem (8) with respect to Ω using graphical

LASSO and let $\hat{\Omega}$ be the resulting estimate of Ω.

3: With $\Omega = \hat{\Omega}$, solve optimization problem (6) with respect to τ. Let $\hat{\tau}$ be the resulting estimate of τ.

4: **if** $|\tau_0 - \hat{\tau}| < e$ and $|(\Omega_0 - \hat{\Omega})ij| < e$ **for** $1 \le i \le j \le p$ **then**

5: Stop and output $(\hat{\tau}, \hat{\Omega})$.

6: **else**

7: Let $\tau_0 = \hat{\tau}, \Omega_0 = \hat{\Omega}$, return to Step 2.

Algorithm 1. Identify the network based on the homogeneous SGGM.

2.3 Heterogeneous SGGM

In the homogeneous SGGM, we assume that a single correlation parameter τ applies to all the subjects. In real data analysis, this parameter may vary across subjects. In this section, we consider network identification without assuming $\tau_1 = \cdots = \tau_m$. Instead, we assume that the parameters τ_i's are independent random variables from the exponential distribution $\tau_i \sim \exp(\alpha)$. Consequently, the joint density function for $\{y_i, \tau_i\}_{i=1}^m$ is

$$\prod_{i=1}^{m} f_i(y_i|\Sigma, \tau_i)\alpha\exp(-\alpha\tau_i) \quad (9)$$

from which the likelihood function for Σ and α is given by

$$L_n(\Omega, \alpha|y) = \int_{\tau_1,\cdots,\tau_m} \prod_{i=1}^{m} f_i(y_i|\Sigma, \tau_i)\alpha_i \exp(-\alpha\tau_i) d\tau_1 \ldots \tau_m \quad (10)$$

With (Equations 9, 10) in hand, the sparse estimate for the network can then be obtained by minimizing the following L_1-penalized negative log-likelihood function:

$$\min_{\Omega,\alpha}\{-2l_n(\Omega, \alpha) + n\lambda|\Omega|_1\} \quad (11)$$

where $l_n(\Omega, \alpha) = \log(L_n(\Omega, \alpha|y))$. Since no explicit form for $l_n(\Omega, \alpha)$ is available, the expectation-maximization (EM) algorithm is proposed here to find the solution to (11) (Dempster et al., 1977). Since we are considering the negative log-likelihood function in (11), the maximization in the EM algorithm will be replaced by the minimization. The correlation parameters $\tau = (\tau_1, \cdots, \tau_m)$ will be taken as the so-called missing data. Recall in the first step of the EM algorithm that the conditional distribution of missing data τ given y, $\Sigma = \Sigma_0$, $\alpha = \alpha_0$ has to be derived from (2) and (9) as follows:

$$\begin{aligned} g(\tau|y, \Sigma_0, \alpha_0) &= \prod_{i=1}^{m} g(\tau_i|y_i, \Sigma_0, \alpha_0) \\ &\propto \prod_{i=1}^{m} |\Phi_i|^{-p/2} \exp(-y_i^T(\Phi_i^{-1} \otimes \Sigma_0^{-1})y_i/2) \exp(-\alpha_0\tau_i) \end{aligned} \quad (12)$$

For (12), the expectation of the complete log-likelihood function for (y, τ) to (12) has to be computed. Given the joint density function (9) of (y, τ), the expectation of its logarithmic transformation can be shown to be

$$\begin{aligned} Q(\Omega, \alpha|\Omega_0, \alpha_0) &= \sum_{i=1}^{m} pE_g\{\log(|\Phi_i|)\} - n\log(|\Omega|) + ntr(S^{(0)}\Omega) \\ &- 2m\log(\alpha) + 2\alpha\sum_{i=1}^{m} E_g\tau_i + n\lambda|\Omega|_1 \end{aligned} \quad (13)$$

where $\bar{S}(\tau) = \frac{1}{n}\sum_{i=1}^{m} S_i(\tau_i)$, $S^{(0)} = E_g\bar{S}(\tau) = \frac{1}{n}\sum_{i=1}^{m} E_g S_i(\tau_i)$ in which $S_i(\cdot)$ is defined in the previous section. In the second step of the EM algorithm, the minimum point of the Q function in (13) has to be computed. This, again, is implemented through a block coordinate descent algorithm. First, for fixed Ω, it is straightforward to show that the minimum of the Q function with respect to α is attained at

$$\hat{\alpha} = \frac{1}{\frac{1}{m}\sum_{i=1}^{m} E_g\tau_i}, \quad (14)$$

i.e., the reciprocal of the sample mean of the conditional expectation of τ_i with respect to density (12). Then, for a given α in (Equation 14), the minimization of (13) with respect to Ω is equivalent to

$$\min_{\Omega}\left\{-\log|\Omega| + tr(S^{(0)}\Omega) + \lambda|\Omega|_1\right\} \quad (15)$$

which can be solved through the graphical LASSO algorithm. The difficult part of this algorithm is to find the expectation $E_g\bar{S}(\tau)$, which may not have an explicit form given that $\bar{S}(\tau)$ is a nonlinear function of τ and the complex form of density function $g(\tau)$; therefore, this expectation will be computed through a Monte Carlo method. This algorithm will be referred to as the heterogeneous longitudinal graphical LASSO, for which the details are summarized in the following table. Given initial values α_0 and Ω_0, tuning parameter λ and error tolerance $e > 0$, the following algorithm is applied.

2.4 Covariate-adjusted SGGM

In the previous section, we assumed that $E\tau_I = 1/\alpha$ is constant across the subjects. In this section, we further relax this constraint, and α can be a function of the covariates. Specifically, we assume $\tau_i \sim exp(\alpha_i)$, where α_i has the following form:

$$\alpha_i = \exp(\alpha_0 + \alpha_1 x_{i1} + \cdots + \alpha_q x_{iq}) = \exp(\alpha^T x_i). \quad (16)$$

1: procedure GIVEN INITIAL VALUES $\boldsymbol{\alpha_0}$ AND $\boldsymbol{\Omega_0}$, TUNING PARAMETER λ AND ERROR TOLERANCE $\boldsymbol{e > 0}$:

2: Sample $\boldsymbol{\tau_{ij}(i = 1, \cdots, m)}$ from the distribution (12) with $\boldsymbol{\Omega = \Omega_0}$;

3: Estimate $\boldsymbol{E(\tau), E_g S_i(\tau_i)}$ and $\boldsymbol{S^{(0)}}$ by $\frac{1}{h}\sum_{j=1}^{h}\boldsymbol{\tau_i}, \frac{1}{h}\sum_{j=1}^{h}\boldsymbol{S_i(\tau_{ij})}$ and $\frac{1}{mh}\sum_{i=1}^{m}\sum_{j=1}^{h}\boldsymbol{S_i(\tau_{ij})}$ respectively

4: Update $\boldsymbol{\alpha}$ by $\hat{\boldsymbol{\alpha}}$ in (14), $\boldsymbol{\Omega}$ by $\hat{\Omega}$, the solution to (15), in which $\boldsymbol{E(\tau)}$ and $\boldsymbol{S^{(0)}}$ are replaced by their estimates in Step 3.

5: **if** $|\boldsymbol{\alpha_0} - \hat{\boldsymbol{\alpha}}| < \boldsymbol{e}$ and $|(\Omega_0 - \hat{\Omega})_{ij}|_1 < \boldsymbol{e}$ **then**

6: Stop and output $(\hat{\boldsymbol{\alpha}}, \hat{\Omega}, \hat{\tau}_i, \boldsymbol{i = 1, \cdots, m})$.

7: **else**

8: $\boldsymbol{\alpha_0} = \hat{\boldsymbol{\alpha}}$ and $\Omega_0 = \hat{\Omega}$, return to Step 2.

Algorithm 2. Identify the network based on the heterogeneous SGGM.

Here, $x_i = (1, x_{i1}, \cdots, x_{iq})^T$ represents covariates such as sex and race, and $\alpha = (\alpha_0, \cdots, \alpha_q)^T$ represents unknown parameters. The model (Equations 9–15) and Algorithm 2 in the previous section can then be revised straightforwardly to accommodate the current regression model (Equation 16). Specifically, first replace the parameter α in (Equations 9, 10) by $exp(\alpha^T x_i)$. Then, the conditional distribution of missing data is given by

$$\begin{aligned} g(\tau|y, \Sigma_0, \alpha_0) &= \prod_{i=1}^{m} g(\tau_i|y_i, \Sigma_0, \alpha_0) \\ &\propto \prod_{i=1}^{m} |\Phi_i|^{-p/2} \exp(-y_i^T(\Phi_i^{-1} \otimes \Sigma_0^{-1})y_i/2) \exp(-\exp(\alpha_0^T x_i)\tau_i) \end{aligned} \quad (17)$$

Based on (Equation 17), the L_1-penalized likelihood estimation of the MIN (11) is then given by

$$\min_{\Omega,\alpha}\{-2l_n(\Omega,\alpha)+n\lambda|\Omega|_1\} \tag{18}$$

For the same reason, we need to use the EM algorithm to solve the optimization problem (Equation 18). The conditional distribution of missing value $\{\tau_i, i=1,\cdots,m\}$ is given by (12), with α_0 replaced by $\exp(\alpha_0^T x_i)$, from which the Q function is given by

$$\begin{aligned}Q(\Omega,\alpha|\Omega_0,\alpha_0) = &\sum_{i=1}^{m} pE_g\{\log(|\Phi_i|)\} - n\log(|\Omega|) + ntr(S^{(0)}\Omega)\\ &- 2\sum_{i=1}^{m}\{\alpha^T x_i - \exp(\alpha^T x_i)E_g\tau_i\} + n\lambda|\Omega|_1,\end{aligned} \tag{19}$$

Given the penalized Q function (Equation 19), the current estimate of the network Ω and parameter α, defined as the minimum point of Q, can be computed by the block coordinate descent algorithm. However, unlike (14), in the current case, the estimate of α does not admit an explicit form. We have to use numerical methods such as Newton algorithms to find the minimizer of the Q function (19). Specifically, we solve the following problem by using the BFGS algorithms:

$$\hat{\alpha} = \min_{\alpha}\left\{\sum_{i=1}^{m}(-\alpha^T x_i + \exp(\alpha^T x_i)E_g\tau_i)\right\} \tag{20}$$

1: **procedure** GIVEN INITIAL VALUES α_0 AND Ω_0, TUNING PARAMETER λ AND ERROR TOLERANCE $e > 0$:

2: Sample $\tau_{ij}(i=1,\cdots,m)$ from the distribution (17);

3: Estimate $E(\tau), E_g S_i(\tau_i)$ and $S^{(0)}$ by $\frac{1}{h}\sum_{j=1}^{h}\tau_i, \frac{1}{h}\sum_{j=1}^{h}S_i(\tau_{ij})$ and $\frac{1}{mh}\sum_{i=1}^{m}\sum_{j=1}^{h}S_i(\tau_{ij})$ respectively

4: Update α by $\hat{\alpha}$ the solution to (20), Ω by $\hat{\Omega}$, the solution to (15), in which $E(\tau)$ and $S^{(0)}$ are replaced by their estimates in Step 3.

5: **if** $|\alpha_0 - \hat{\alpha}| < e$ and $|(\Omega_0 - \hat{\Omega})_{ij}|_1 < e$ **then**

6: Stop and output $(\hat{\alpha}, \hat{\Omega}, \hat{\tau}_i, i=1,\cdots,m)$.

7: **else**

8: $\alpha_0 = \hat{\alpha}$ and $\Omega_0 = \hat{\Omega}$, return to Step 2.

Algorithm 3. Identify the network based on the covariate-adjusted SGGM.

Using $\hat{\alpha}$ in (Equation 20), the estimate of the MIN Ω is given by the solution of Equation (15) through the graphical LASSO algorithm. The algorithm is summarized in the following.

Remark: (1) All three algorithms described above leverage the graphical LASSO algorithm to achieve efficiency even though graphical LASSO itself is devised for independent data. (2) Note that correlation τ_i for subject i is a random variable. The forecast of τ_i, $\hat{\tau}_i$ is given by the expectation of distribution (12), which is one of the outputs in Algorithms 2 and 3. (3) The proposed algorithms can generate a solution path for a given sequence of tuning parameters λ. To select the optimal network Ω from the candidate networks, model selection criteria can be used, e.g., Akaike information criterion (AIC), Bayesian information criterion (BIC), or cross-validation (CV). In the numerical studies in the next section, we use the extended BIC (EBIC) that is dedicated to the graphical model to select λ (Foygel and Drton, 2010).

3 Results

3.1 Simulation

In this section, we compare the proposed algorithms, which are referred to as longitudinal graphical LASSO (LGLASSO) algorithms, with other existing network selection methods. These methods include graphical LASSO (Friedman et al., 2008; Friedman et al., 2019), neighborhood (Meinshansen and Bühlmann, 2006), GGMselectC01 and GGMselect-LA algorithms (Giraud et al., 2012; Bouvier et al., 2022). The graphical LASSO and neighborhood algorithms have been used in the SPIEC-EASI pipeline in (Kurtz et al., 2015) to select MINs and both are based on the L_1-penalty. On the other hand, the GGMselect algorithm family provides different ways to construct and select the candidate models, e.g., GGMselectC01 employs the estimation procedure in (Wille and Bühlmann, 2006) to construct the candidate models, while GGMselect-LA uses the Fisher random variable to define the criterion for network selection. We demonstrate that the proposed longitudinal graphical LASSO algorithms can outperform these existing algorithms for simulated high-dimensional longitudinal microbiome data.

Let TP,P,FP,N,TN be the numbers of true positive edges, real positive edges, false-positive edges, null edges, and true null edges, respectively. In the following, we use the true/false positive rate (TPR/FPR) to measure the performance of each algorithm. They are defined as TPR = TP/P, FPR = FP/N. Also, for the conventional indices sensitivity/specificity, we have sensitivity = TPR, specificity = TN/N

The simulations consist of three scenarios. In the first scenario, we consider cases where the data follow the SGGM in Sections 2.2 and 2.3. Recall that the SGGM assumes that all the taxa in the microbiome share a common dampening rate τ_i for subject i. We refer to such microbiomes as having a homogeneous community. In the second scenario, we consider the microbiomes that violate such homogeneity, i.e., having heterogeneous communities. In scenario three, we consider the left-censored microbiome data, which aims to test the robustness of the SGGM with respect to zero inflation. The zero-inflation phenomena are widely observed in 16S rRNA gene sequencing experiments and violate the assumptions of the SGGM. Both the homogeneous and heterogeneous versions of LGLASSO in Sections 2.2 and 2.3 are investigated in each scenario. We use the receiver operating characteristic (ROC) curve to show the superiority of our algorithms over the graphical LASSO and the neighborhood algorithms in all scenarios. In the case of the heterogeneous LGLASSO, we also use (TPR,FPR) to compare the performances of the algorithms in which the networks

are selected through the extended BIC (EBIC). The methods of GGMselect-C01 and GGNselect-LA have their own model selection approaches for network selection. The EBIC for the Gaussian graphic model is given by

$$\mathbf{EBIC}(\boldsymbol{G}) = -2\boldsymbol{l_n}(\boldsymbol{G}) + |\boldsymbol{G}|\mathbf{log}(\boldsymbol{n}) + |\boldsymbol{G}|\mathbf{log}(\boldsymbol{p})/\boldsymbol{T} \quad (21)$$

where l_n is the log-likelihood function, n is the sample size, G is the network of interest, $|G|$ is the number of edges in G, p is the number of nodes, and T is the tuning parameter. In (Equation 21), we choose a typical value $T = 2$ for model selection.

3.1.1 Scenario 1: homogeneous microbial community

Since homogeneity/heterogeneity involves both microbial community and LGLASSO algorithms, we will use more specific names, homogeneous-subject LGLASSO (heterogenoussubject LGLASSO), to refer to the algorithm to avoid confusion in the following sections.

In a homogeneous microbial community, for each subject, all taxa in the microbiome share a common correlation-dampening rate. Based on whether the subjects share dampening rates, we have the homogeneous-subject LGLASSO in Section 2.2 and heterogeneous-subject LGLASSO in Sections 2.3 and 2.4, respectively. We first consider the former case. Specifically, we consider networks with $p = 80$ nodes shared by all $m = 10$ subjects. The precision matrix corresponding to this network is generated through the R package *BDgraph* (Mohammadi and Wit, 2019) with an edge density equal to 0.1. For subject i ($1 \le i \le m$), there are n_i observations where n_i follows Poisson distribution with the mean value of 10. The spaces between two consecutive time points t_{ij} and $t_{i(j+1)}$ are generated by $\max\{\Delta_{ij}, 0.5\}$, where Δ_{ij} follows a Poisson distribution with a mean value of 1. Then, the data for subject i are generated with the real autocorrelation parameter in Algorithm 1 given by τ.

First, the homogeneous-subject LGLASSO, graphical LASSO, and neighborhood algorithms are carried out for the simulated data set, from which their respective solution paths are computed. For each path, the connection probability p_{ij} for any given node pair (V_i, V_j) is then computed as the proportion of networks that have (V_i, V_j) connected among all the networks on the path. The processes are replicated 50 times, and the final estimate for p_{ij} is the average of these 50 estimates of p_{ij}. With these p_{ij} ($1 \le i \le j \le 80$) in hand, the ROC curve is computed based on the R package *pROC* (Robin et al., 2011) and displayed in Figure 1 under the two indicated situations $\tau = 0.14, 0.018$. In both cases, the proposed homogeneous-subject LGLASSO algorithm outperforms the graphical LASSO and the neighborhood algorithms. The differences between these ROC curves are more evident for $\tau = 0.018$ than $\tau = 0.14$. We interpret this phenomenon as the proposed LGLASSO can better handle the correlated longitudinal data than other algorithms.

Next, we consider the heterogeneous-subject LGLASSO and covariate-adjusted LGLASSO algorithms proposed in Sections 2.3 and 2.4, respectively. For ease of exposition, we consider the heterogeneous-subject LGLASSO as a special covariateadjusted LGLASSO in which vector 1 is the only covariate. As in Scenario 1, we still consider a network with $p = 80$ nodes and an edge density equal to 0.1. For each subject of $m = 10$ subjects, the spaces between consecutive observations are generated similarly. Since the dampening rate τ is an exponential random variable in the heterogeneous-subject LGLASSO, we generate random samples as the individual τ_is from the following two settings, $E\tau = 0.14, 0.018$. Then with the same replication scheme as above, the ROC curves are computed and plotted in Figure 2. Similar to the case of homogeneous-subject LGLASSO, these ROC curves also demonstrate the superiority of heterogeneous-subject LGLASSO over other methods, especially for the higher correlation case $E\tau = 0.018$.

We then employ the EBIC (21) to select the optimal model from the solution path. Specifically, two covariates x_1 and x_2 are introduced in which $x_1 \sim N(0,1)$ and $P(x_2 = 0) = P(x_2 = 1) = 0.5$. The three settings for their coefficients (α_1, α_2) in Equation (16) are $(\alpha_1, \alpha_2) = (0,0), (0.5, 0.5), (1,1)$ while the intercept is always $\alpha_0 = 4$. Note that with $(\alpha_1, \alpha_2) = (0,0)$, the covariates have no effect on the dampening rate, and therefore, the model reduces to the heterogeneous-subject LGLASSO in Section 2.3. For each

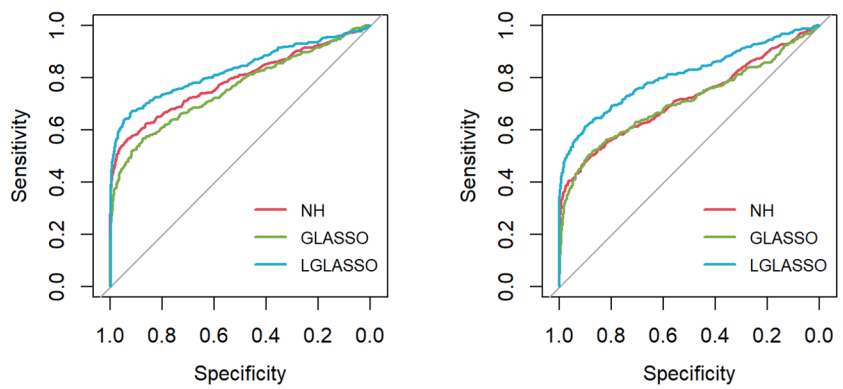

FIGURE 1
ROC curves for the homogeneous-subject LGLASSO, graphical LASSO, and neighborhood algorithms. The data are generated based on the homogeneous SGGM in Section 2.2. The dampening rates for the left and right plots are $\tau = 0.36, 0.018$, respectively.

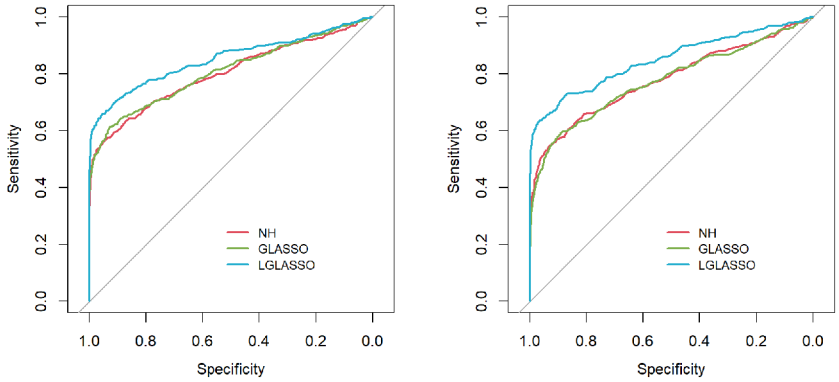

FIGURE 2
ROC curves for the heterogeneous-subject LGLASSO, graphical LASSO, and neighborhood algorithms. The data are generated based on the heterogeneous SGGM in Section 2.3. The average dampening rates for the left and right plots are $E\tau$ = 0.14,0.018, respectively.

simulated data set, the pairs (TPR,FPR) can be computed. The process is replicated 50 times and the averages of these 50 (TPR, FPR) are listed in Table 1. Note that for GGMselect-C01 and GGMselect-LA listed in Table 1, we used their own model selection method instead of the EBIC. From Table 1, we can see that the proposed LGLASSO algorithm obtains the highest TPR and lowest FPR in most cases. In other words, with the EBIC as the model selection method, the heterogeneous-subject LGLASSO algorithms still have the best performance among the algorithms considered.

3.1.2 Scenario 2: heterogeneous microbial community

In Section 2.1, we mentioned that the taxa from the same subject are supposed to share the same correlation-dampening rate. In this section, we show by simulation that even if the microbiome fails to satisfy this assumption, the proposed algorithm can still outperform the conventional methods. Specifically, in the heterogeneous microbial community, the taxa have different correlation-dampening rates for each subject. For ease of exposition, we consider a simple situation where the microbiome

TABLE 1 Performance comparison of heterogeneous-subject LGLASSO (LGLASSO), graphical LASSO (GLASSO), neighborhood (NH) algorithm, GGMselect-C01 (C01), and GGMselect-LA (LA).

			α = (0,0)		α = (0.5,0.5)		α = (1,1)	
			TPR	FPR	TPR	FPR	TPR	FPR
m = 10	$E(n_i)$ = 5	GLASSO	0.489	0.282	0.508	0.299	0.480	0.311
		NH	0.522	0.297	0.567	0.311	0.541	0.327
		C01	0.138	0.037	0.295	0.126	0.321	0.174
		LA	0.328	0.111	0.369	0.125	0.367	0.151
		LGLASSO	**0.603**	**0.276**	**0.617**	**0.274**	**0.591**	**0.256**
m = 10	$E(n_i)$ = 10	GLASSO	0.545	0.327	0.551	0.314	0.476	0.318
		NH	0.654	0.394	0.640	0.371	0.568	0.339
		C01	0.528	0.380	0.509	0.320	0.458	0.310
		LA	0.583	0.291	0.549	0.252	0.546	0.289
		LGLASSO	**0.690**	**0.246**	**0.671**	**0.220**	**0.586**	**0.203**
m = 10	$E(n_i)$ = 20	GLASSO	0.631	0.370	0.576	0.329	0.534	0.316
		NH	0.737	0.506	0.687	0.412	0.644	0.374
		C01	0.852	0.801	0.809	0.744	0.820	0.748
		LA	0.746	0.505	0.730	0.473	0.711	0.448
		LGLASSO	**0.722**	**0.213**	**0.732**	**0.266**	**0.696**	**0.239**

The data are generated based on the heterogeneous SGGM in Section 2.3.
The bold values are the results of the proposed algorithm LGLASSO.

consists of two subcommunities, A and B, which have different correlation-dampening rates τ_A and τ_B, respectively. Furthermore, we assume that these two communities are independent of each other, i.e., the taxa in community A (B) can correlate with one another; however, they are independent of the taxa in community B (A).

Specifically, for $p = 80$ taxa, we assume the first 40 taxa are in community A, and the other 40 taxa are in community B. As in Scenario 1, we first investigate the performance of the homogeneous-subject LGLASSO algorithm. For $m = 10$, $En_i = 10$, $\tau_A = 0.018$, with a given τ_B, the data for communities A and B are generated by the same method as that in Scenario 1. Since we assume these two communities are independent, the complete data set is simply a combination of the data sets from communities A and B. Three solution paths for the homogeneous-subject GLASSO, graphical LASSO, and neighborhood algorithms are then computed based on the combined data set from which the estimates of the connection probability p_{ij}s are computed the same way as in Scenario 1. Based on these p_{ij}s, the ROC curves are plotted in Figure 3 for the three settings, $\tau_B = 0.36$,0.049, and 0.018. Obviously, for heterogenous communities, the proposed homogeneous-subject LGLASSO still outperforms the other two methods, especially when the correlation is higher (i.e., $\tau_B = 0.018$)

Next, we consider the performance of the heterogeneous-subject LGLASSO. The data are generated from heterogeneous subjects with heterogeneous microbial communities. For ease of exposition, we focus on the heterogeneous-subject model in Section 2.3, and the covariate-adjusted model in Section 2.4 is omitted here. Specifically, for subject i, the corresponding microbiome consists of two microbial communities that have correlation dampening rates τ_{iA} and τ_{iB} $(1 \le i \le m)$ and satisfy $\tau_{iA} \sim \exp(\alpha 1)$, $\tau_{iB} \sim \exp(\alpha_2)$. The parameter settings for $(E\tau_{iA}, E\tau_{iB})$ include the following three cases: $(E\tau_{iA}, E\tau_{iB})$ = (0.036,0.36),(0.036,0.14),(0.036,0.049). For each (α_1, α_2) pair, 10 random samples are generated for (τ_{iA}, τ_{iB}) from the corresponding exponential distributions, which are used as the real autocorrelation parameters for the 10 subjects. With (τ_{iA}, τ_{iB}), n_i measurements for subject i are then generated in the same way as in Figure 3 with $En_i = 10$. With these data, the heterogeneous-subject LGLASSO, graphical LASSO, and neighborhood algorithms are carried out, and the resulting ROC curves are presented in Figure 4. These ROC curves demonstrate that with heterogeneous communities and heterogeneous subjects, the proposed algorithm LGLASSO still outperforms graphical LASSO and neighborhood methods when the correlations between data are reasonably high.

3.1.3 Scenario 3: zero-inflated relative abundances

In this section, we consider the performance of the algorithms when the data generated from the SGGM are left-censored. Left-censored data represent the transformed zero-inflated relative abundance of 16S rRNA gene sequences. An example of such transformations is shown in Section 3.2 in the real data analysis. Here, we investigate the influence of zero inflation on the performance of the proposed algorithms. As in Scenario 1, we consider left-censored homogeneous-subject LGLASSO with $m = 10$, $En_i = 10$ $(1 \le i \le m)$ and $\tau = 0.018$. Under this setting, the data are first generated in the same way as in Scenario 1. Using the generated data, we consider the following censoring scheme: given $0 < q_1 < 1$, for each taxon, all observations with values less than quantile y_{q1} are replaced by y_{q1} with a probability of q_2. This censoring scheme is motivated by the observation that the smaller the relative abundance is, the higher the probability is that this taxon is missed by the sequencing experiments. Here, we consider six combinations of (q_1, q_2), i.e., (0.1,0.3),(0.1,0.5),(0.1,0.7),(0.4,0.3), (0.4,0.5),(0.4,0.7). For each of these combinations, the ROC curves corresponding to the respective solution paths of the homogeneous-subject LGLASSO, graphical LASSO, and the neighborhood algorithms are shown in Figure 5. Obviously, even though the proposed homogeneous-subject LGLASSO algorithm outperforms the other algorithms, zero inflation can significantly affect its performance, and the advantage of LGLASSO diminishes when the proportion of zero is high. The same investigations are carried out for the heterogeneous-subject LGLASSO. The procedure is the same as the above homogeneous case except that the dampening

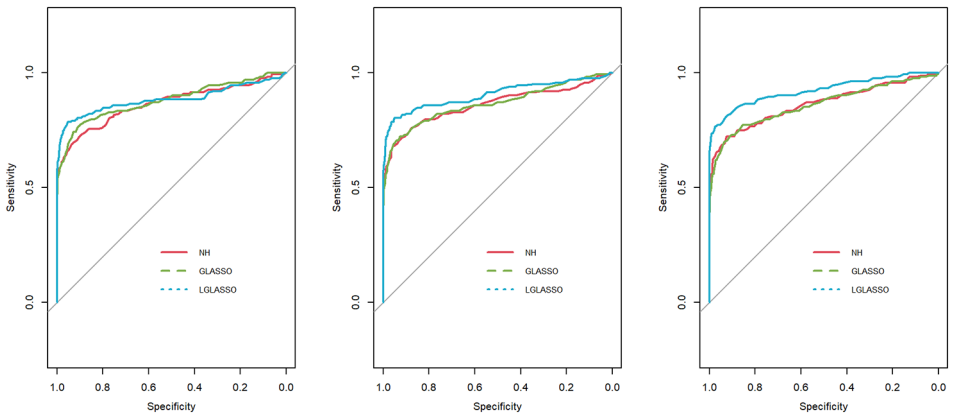

FIGURE 3
ROC curves for the homogeneous-subject LGLASSO, graphical LASSO, and neighborhood algorithms. The data are generated from homogeneous SGGM with a heterogeneous microbiome. The dampening rates, from left to right, are (τ_A, τ_B) = (0.018,0.36), (0.018,0.049), and (0.018,0.018), respectively.

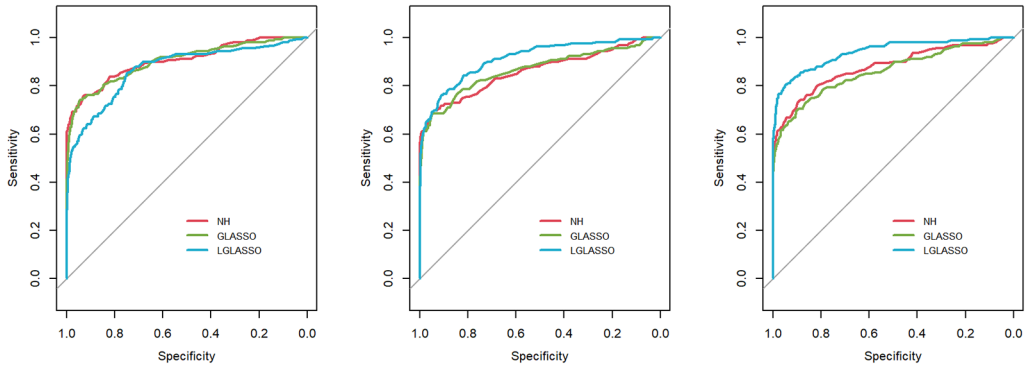

FIGURE 4
ROC curves for the heterogeneous-subject LGLASSO, graphical LASSO, and neighborhood algorithms. The data are generated from heterogeneous SGGM with a heterogeneous microbiome. The average dampening rates, from left to right, are $(E\tau_{iA}, E\tau_{iB})$ = (0.036,0.36), (0.036,0.14), and (0.036,0.049), respectively.

rate is random with $E\tau_i = 0.018$ for $(1 \leq i \leq m)$. The resulting ROC curves are depicted in Figure 6, from which we can see a similar pattern as the ones in Figure 5.

3.2 Gut microbial interaction network and phylogenetic tree

In this section, a longitudinal data set from a cohort of children with cystic fibrosis was investigated using the homogeneous version of the proposed algorithm in Section 2.2. Specifically, stool samples from thirty-eight children were collected from children aged 6 months to 51 months old (Madan et al., 2012). The number of observations from each child ranged from 2 to 17. Each observation consisted of the abundance of 16,383 amplicon sequence variants (ASVs) of the 16S rRNA gene. These sequences were then collapsed to the genus level using the R package *DADA2* (Callahan et al., 2016). The sequences that had no genus-level information were dropped. Then, all the taxa with a proportion of nonzero observations less than 10% were combined, which was referred to as the composite taxon. There were 83 total remaining taxa. The observations of zeros for each of these 83 taxa were replaced by the

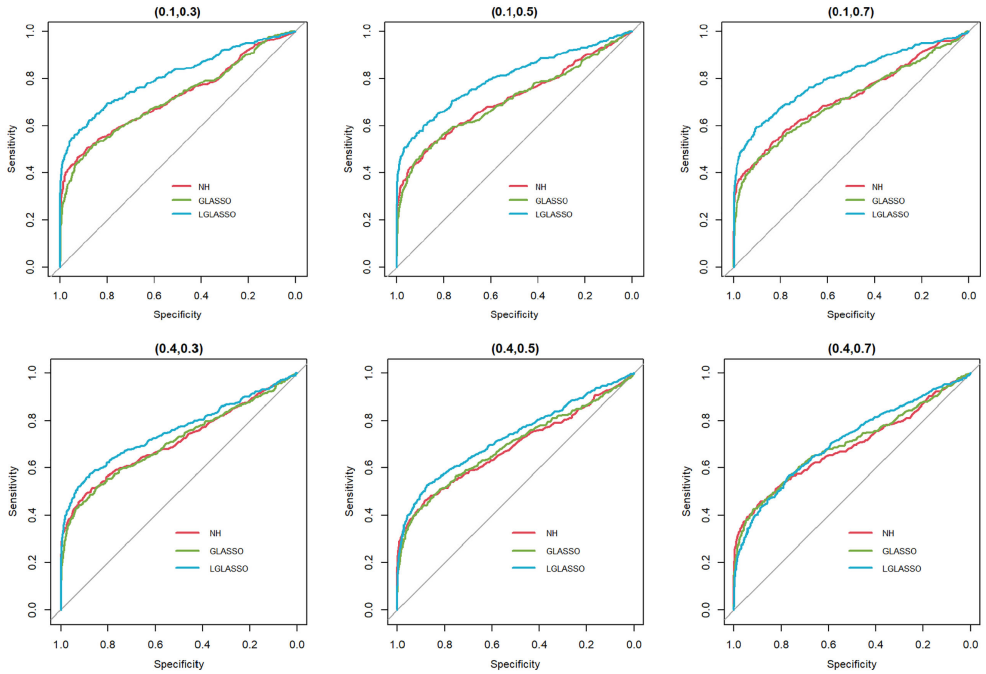

FIGURE 5
ROC curves for the homogeneous-subject LGLASSO, graphical LASSO, and neighborhood algorithms. The data are generated from a left-censored homogeneous SGGM model. The first number in parentheses is the quantile, and the second is the probability, which are the parameters for data censoring.

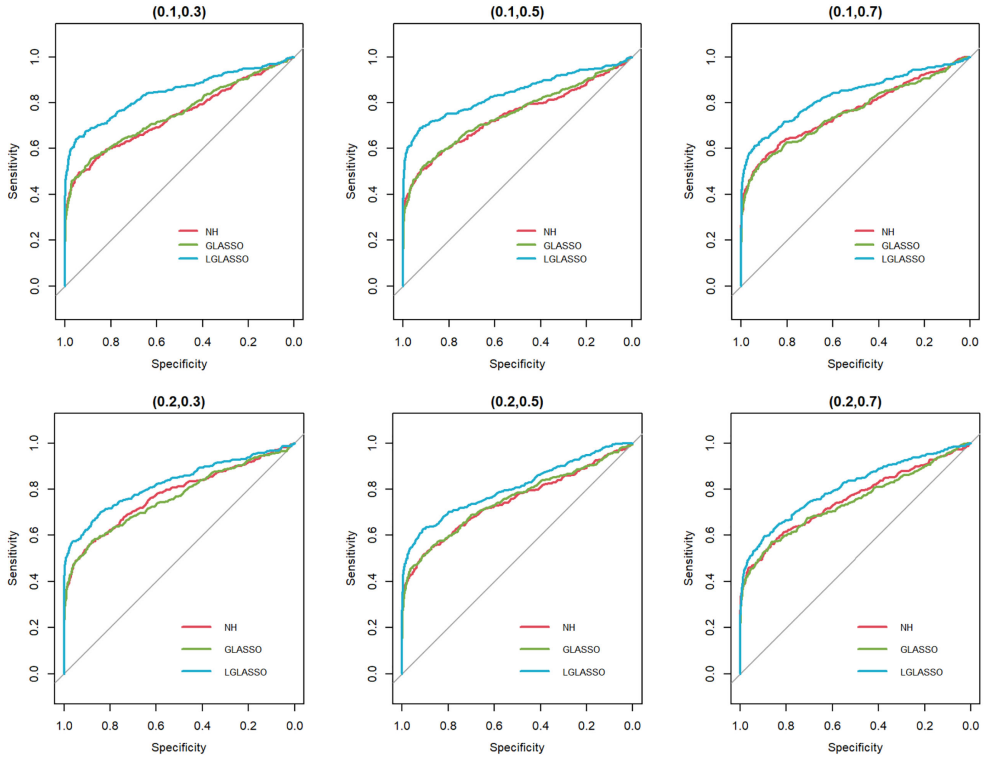

FIGURE 6
ROC curves for the heterogeneous-subject LGLASSO, graphical LASSO, and neighborhood algorithms. The data are generated from a left-censored heterogeneous SGGM model. The first number in parentheses is the quantile, and the second is the probability, which are the parameters for data censoring.

minimum abundance of that taxon divided by 10. The log-ratio transformation was then carried out to obtain the relative abundance, in which the composite taxon was used as the reference. Similar log-ratio transformations have been used and justified in empirical studies (Kurtz et al., 2015; Greenacre et al., 2021).

The application of the homogeneous model in Section 2.2 to the transformed data yielded the estimated network, which is displayed in Figure 7. Based on the modularity maximization algorithm (Newman, 2006; Blondel et al., 2008), five communities were identified in the estimated network, which is listed in Table 2.

To show that the estimated network can reveal the true structure of the underlying network, we investigated the relationship between the estimated network and the phylogenetic tree of the 82 taxa (the composite network was excluded here). The phylogenetic tree constructed from the same data set is presented in Figure 8 and demonstrates the evolutionary relationship among these taxa. Our underlying hypothesis is that microbial taxa that are closer in terms of evolutionary history also have more/stronger interactions in the human body. To validate this hypothesis, the null hypothesis is set as follows: the estimated network in Figure 7 is independent of the phylogenetic tree in Figure 8. Correlation between the estimated network and phylogenetic tree is employed to test this hypothesis. In particular, we computed the distances between two taxa in the estimated network and the phylogenetic tree. Here, the distance between taxa A and B is defined as the length of the shortest path from A to B in the estimated network (phylogenetic tree). If no paths exist between two taxa in the network (phylogenetic tree), that pair will be excluded from the following computations. Let d_1 and d_2 be the distance vectors for all possible pairs of taxa from the estimated network and phylogenetic tree, respectively. The correlation between d_1 and d_2 is used to measure the relatedness between the network and phylogenetic tree. This correlation was determined to be $r_0 = 0.333$ for the tuning parameter selected by the EBIC.

To understand the significance of $r0$ against the null hypothesis, we use the permutation method to estimate the null distribution. Specifically, we keep the structure of the estimated network unchanged and permute the order of the 82 taxa $m = 5000$ times on the estimated network. *Let* $d_1^{(i)} (i = 1, \cdots, 5000)$ be the distance vectors of the network for the ith permutation. Then, the correlations $r^{(i)} = cor(d_1^{(i)}, d_2), i = 1, \cdots, 5000$, which collectively depict the null distribution, can be derived. Given $r^{(i)}$, the p value of r_0 is smaller than 1/5000, which means that the correlation between the estimated network and phylogenetic tree is statistically significant, i.e., the data support the hypothesis that microbial taxa with closer evolution histories tend to have more/stronger interactions.

It should be noted that some related findings have been reported in the literature. In (Chaffron et al., 2010), the authors performed a global meta-analysis of previously sampled microbial lineages in the environment. They found that genomes from

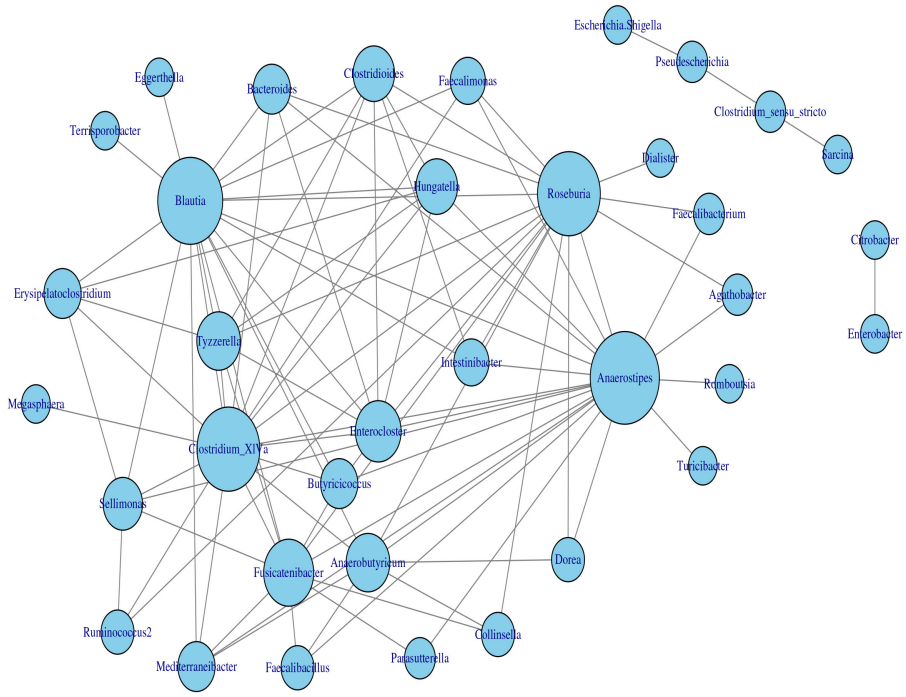

FIGURE 7
Microbial interaction network generated with the homogeneous LGLASSO based on the gut microbiome abundance data in Section 3.2.

coexisting taxa tended to be more similar than expected by chance, both with respect to pathway content and genome size. The studies in (Eiler et al., 2012) also revealed that ecological coherence is often dependent on taxonomic relatedness. These studies employed coefficient-based methods such as Fisher's exact test to infer the interaction of taxa. This can lead to a misleading conclusion. It is known that the correlation between two taxa, A and B, may be induced by their correlation with a third taxon C, even though A and B are independent if C is fixed. In the current study, the interaction between the taxa is defined based on the conditional correlation coefficient, which by its definition eliminates the possible spurious correlation between taxa A and B induced by taxon C. Therefore, by using abundance instead of cooccurrence information and boosted by the proposed methods, we reach a more convincing and robust conclusion than existing ones in the literature.

TABLE 2 Five communities selected by maximizing the modularity of the estimated MIN in Figure 7.

C1	Escherichia.Shigella, Clostridium.sensu.stricto, Sarcina, Pseudescherichia
C2	Blautia, Erysipelatoclostridium, Bacteroides, Tyzzerella, Megasphaera, Intestinibacter, Enterocloster, Hungatella, Terrisporobacter, Clostridioides, Clostridium.XlVa, Butyricicoccus
C3	Anaerostipes, Fusicatenibacter, Agathobacter, Faecalibacterium, Dorea, Collinsella, Faecalimonas, Mediterraneibacter, Romboutsia, Faecalibacillus, Anaerobutyricum, Roseb Parasutterella, Dialister, Turicibacter
C4	Enterobacter, Citrobacter
C5	Sellimonas, Ruminococcus2

The results above are derived with the fixed tuning parameter selected by the EBIC. The correlations between the estimated MIN and phylogenetic tree are actually robust with respect to the tuning parameter. To demonstrate this point, let us consider the networks generated from other tuning parameter choices. Specifically, for each of the 40 tuning parameters ranging from $\lambda = 7$ to 13, the same homogeneous model and the permutation test are carried out. The estimated and permuted correlations are displayed in Figure 9 in the form of a boxplot. From left to right, the boxplots in Figure 9 correspond to the tuning parameters increasing from 7 to 13. The dots linked by the line represent the correlations between the estimated MINs and the phylogenetic tree, while others correspond to the correlations computed from the permuted MINs. The most prominent feature of Figure 9 is that for all 40 cases, the observed correlations between the MINs and the phylogenetic tree are positive; in the first half of the boxplots, the estimated correlations are also significant. It should be noted that even though the estimated correlations in the second half of the boxplots do not appear significant, it does not mean that the edges in these estimated networks do not reflect the true structure. Instead, the insignificance may stem from the fact that the second half of the networks are sparser. A sparse network will generate shorter vectors d_1, d_2, which in turn increase the variability of the correlation estimates.

Finally, let us compare the results of the proposed algorithm with those of the SPIECEASI algorithm in (Kurtz et al., 2015). Though in the original form of the SPIEC-EASI algorithm, the centered log-ratio transformation was employed for the relative abundances of the taxa, we use the additive log-ratio transformation here. Note that the SPIEC-EASI algorithm includes both graphical LASSO and neighborhood algorithms. We restrict ourselves to networks with

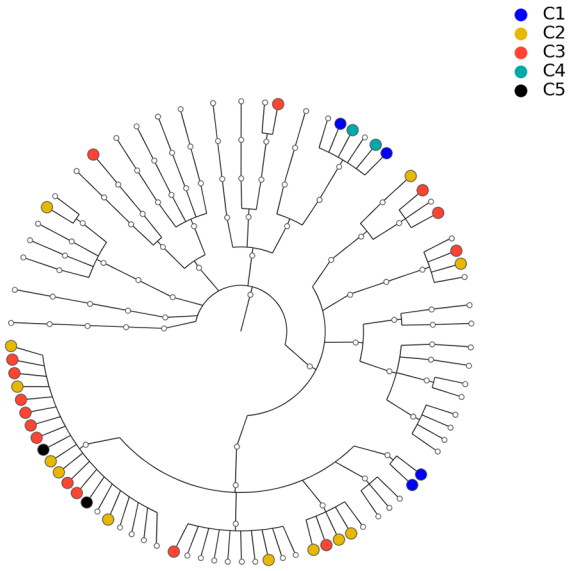

FIGURE 8
Phylogenetic tree for the 82 microbial taxa in Section 3.2. Dots with different colors correspond to different communities listed in Table 2.

edges between 20 and 1300, which we believe cover all biologically meaningful cases. First, we compute the solution path for each algorithm. For the networks on each solution path, their correlations with the phylogenetic tree are calculated. Figure 10 displays the correlations for the three solution paths corresponding to the three algorithms. It is evident that for most of the solution paths, the networks generated through the proposed algorithm have higher correlations with the phylogenetic tree than those generated through the SPIEC-EASI algorithms. The correlations at the beginning parts of the three paths appear to be comparable. We attribute this to the fact that the networks at the beginning parts are much sparser, leading to a smaller sample size when computing the correlations. A smaller sample size can blur the comparison of different algorithms, as shown in Figure 9. In other words, if we assume that the phylogenetic tree represents the true structure of the MIN, then the proposed algorithms have greater power in the identification of the MIN than that of the SPIEC-EASI algorithm.

4 Discussion

Identifying microbial interaction networks is critical for understanding the causal relationship among taxa. However, it remains a challenging problem since observations of the microbiome have many distinct features, such as high dimensionality, zero inflation, and composition. In this study, we study network identification based on irregularly spaced longitudinal 16S rRNA gene sequencing data. For microbial abundance data, the correlations between different time points are typically omitted in practice due to technical difficulties. In this study, a model named SGGM is proposed to characterize the correlations in the longitudinal microbial abundance data. Efficient inference algorithms for estimating microbial interaction networks are devised based on the SGGM. Through the use of simulated data, our model and algorithms show that they have more power to identify microbial interaction networks than conventional methods, where the correlations are just omitted. Furthermore, the algorithms demonstrate their robustness when the data do not follow the SGGM strictly, e.g., heterogeneous microbial communities and zero inflation. The proposed method is employed to study the microbiomes from a cohort with cystic fibrosis disease. The relationship between the microbial interaction networks and the phylogenetic tree is revealed, strengthening previous literature results. It is also necessary to highlight the limitations of SGGM and the related LGLASSO algorithms. First, SGGM only models the stationary process, i.e., the microbial correlation structure remains the same during the data collection process. This may or may not be a valid assumption for a specific situation. For example, the subjects may get vaccinated during the data collection period, which may affect how the constituent microbes interact with each other. If this is the case, SGGM should not be used. Second, SGGM assumes a constant dampening rate τ for all the

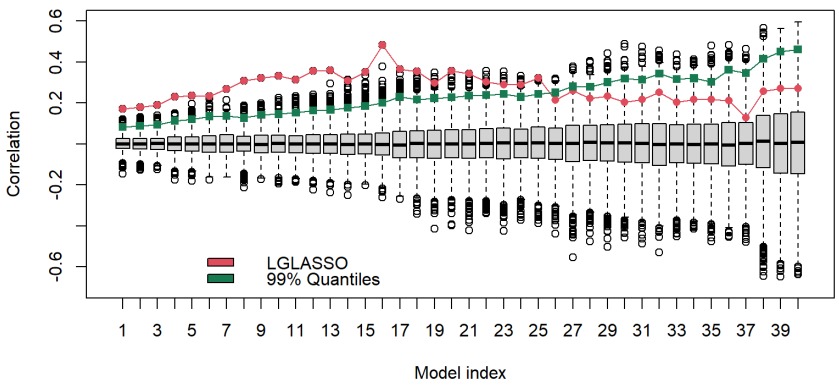

FIGURE 9
Boxplots of correlations between the networks and phylogenetic tree that correspond to, from left to right, the tuning parameters λ increasing from 7 to 13. For a given boxplot, the red dot represents the correlations from the original estimated networks, and the black dots represent correlations computed from 5000 permutations of the estimated network. The green dot represents the 1% upper quantile of the 5000 permutation correlations.

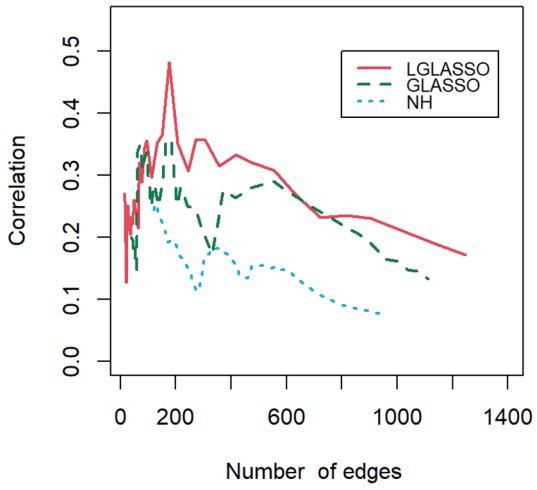

FIGURE 10
Correlations between the phylogenetic tree and networks selected by the LGLASSO, GLASSO, and NH methods for a sequence of tuning parameters.

taxa in the microbiome within the same subject. We only studied the robustness of the LGLASS algorithms with respect to τ under very simple cases, i.e., two independent sub-microbial communities with different dampening rates. In reality, things can get very involved. For example, the whole community may have multiple sub-communities, and each of them has its own dampening rate, e.g., community A evolves with a high frequency, community B evolves with a medium frequency, community C evolves with a low frequency, and communities A, B, and C are related to each other in some way. In such cases, it should be cautious to use SGGM to identify the underlying network. Some form of cross-validation is recommended in such situations.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author. The proposed algorithms are available in the R package lglasso, which can be accessed at https://github.com/jiezhou-2/lglasso. The code for reproducing the simulation results in Section 3.1 can be found at https://jiezhou-2.github.io/lglasso_data_analysis/index.html. For access to the cystic fibrosis data used in Section 3.2, please contact the corresponding author at Anne.G.Hoen@dartmouth.edu.

Author contributions

JZ: Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. JG: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Supervision, Writing – review & editing. WV: Conceptualization, Methodology, Writing – review & editing. HC: Investigation, Writing – review & editing. SL: Writing – review & editing. JM: Data curation, Funding acquisition, Project administration, Writing – review & editing. MC: Conceptualization, Resources, Writing – review & editing. AH: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work is supported by the NIH NIGMS grants P20GM125498, P20GM130454, R01LM012723 and NIAID grant U19AI145825.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., et al. (2011). Enterotypes of the human gut microbiome. *Nature* 473, 174–180. doi: 10.1038/nature09944

Avella-Medina, M., Battey, H. S., Fan, J., and Li, Q. (2018). Robust estimation of high-dimensional covariance and precision matrices. *Biometrika* 105, 271–284. doi: 10.1093/biomet/asy011

Bach, F. R., and Jordan, M. I. (2004). Learning graphical models for stationary time series. *IEEE Trans. Signal process.* 52, 2189–2199. doi: 10.1109/TSP.2004.831032

Barberan, A., Bates, S. T., Casamayor, E. O., and Fierer, N. (2012). Using network analysis to explore co-occurrence patterns in soil microbial communities. *ISME J.* 6, 343–351. doi: 10.1038/ismej.2011.119

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.* 2008 (10), P10008. doi: 10.1088/1742-5468/2008/10/P10008

Bouvier, A., Giraud, C., Huet, S., and Verzelen, N. *GGMselect: Gaussian Graphs Models Selection, 2022, version: 0.1-12.5* (CRAN). Available at: https://CRAN.R-project.org/package=GGMselect.

Callahan, B. J., McMurdie, P. J., Rosen, M. L., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13, 581–583. doi: 10.1038/nmeth.3869

Chaffron, S., Rehrauer, H., Pernthaler, J., and Von Mering, C. (2010). A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res.* 20, 947–959. doi: 10.1101/gr.104521.109

Chen, I., Kelkar, Y. D., Gu, Y., Zhou, J., Qiu, X., and Wu, H. L. (2017). Highdimensional linear state space models for dynamic microbial interaction networks. *PloS One* 12 (11), e0187822. doi: 10.1371/journal.pone.0187822

Chung, H. C., Gaynanova, I., and Ni, Y. (2022). Phylogenetically informed Bayesian truncated copula graphical models for microbial association networks. *Ann. Appl. Stat.* 16 (4), 2437–2457. doi: 10.1214/21-AOAS15

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R Stat. Soc. Ser. B Methodol.* 39, 1–22. doi: 10.1111/j.2517-6161.1977.tb01600.x

Diggle, P., Diggle, D., Diggle, P. J., Heagerty, P., Liang, K.-Y., Zeger, S., et al (2002). *Analysis of Longitudinal Data* (London: Oxford University Press). doi: 10.1093/oso/9780198524847.001.0001

Eiler, A., Heinrich, F., and Bertilsson, S. (2012). Coherent dynamics and association networks among lake bacterioplankton taxa. *ISME J.* 6, 330–342. doi: 10.1038/ismej.2011.113

Epskamp, S., Waldorp, L. J., Mõttus, R., and Borsboom, D. (2018). The gaussian graphical model in cross-sectional and time-series data. *Multivariate Behav. Res.* 53, 453–480. doi: 10.1080/00273171.2018.1454823

Fang, H., Huang, C., Zhao, H., and Deng, M. (2017). gCoda: conditional dependence network inference for compositional data. *J. Comput. Biol.* 24, 699–708. doi: 10.1089/cmb.2017.0054

Faust, K., and Raes, J. (2012). Microbial interactions: from networks to models. *Nat. Rev. Microbiol.* 10, 538–550. doi: 10.1038/nrmicro2832

Foygel, R., and Drton, M. (2010). Extended Bayesian Information Criteria for Gaussian Graphical Models. In: *Advances in Neural Information Processing Systems* (Curran Associates, Inc). Available online at: https://papers.nips.cc/paper/2010/hash/072b030ba126b2f4b2374f342be9ed44-Abstract.html (Accessed November 21, 2021).

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical LASSO. *Biostatistcs* 9, 432–441. doi: 10.1093/biostatistics/kxm045

Friedman, J., Hastie, T., and Tibshirani, R. (2019). *Graphical Lasso: Estimation of Gaussian Graphical Models Version: 1.11*. Available online at: https://CRAN.R-project.org/package=glasso.

Gause, G. F. (1934). *The Struggle for Existence* (New York: Williams and Wilkins). doi: 10.5962/bhl.title.4489

Giraud, C., Hue, S., and Verzelen, N. (2012). Graph selection with GGMselect. *Stat. Appl. Genet. Mol. Biol.* 11 (3). doi: 10.1515/1544-6115.1625

Greenacre, M., Martínez-Alvaró, M., and Blasco, A. (2021). Compositional data analysis of microbiome and any-omics datasets: A validation of the additive logratio transformation. *Front. Microbiol.* 12. doi: 10.3389/fmicb.2021.727398

Harcombe, W. (2010). Novel cooperation experimentally evolved between species. *Evolution* 64, 2166–2172. doi: 10.1111/j.1558-5646.2010.00959.x

He, L., Wang, C., Hu, J., Gao, Z., Falcone, E., Holland, S. M., et al. (2022). ARZIMM: A novel analytic platform for the inference of microbial interactions and community stability from longitudinal microbiome study. *Front. Genet.* 13. doi: 10.3389/fgene.2022.777877

Jiang, D., Sharpton, T., and Jiang, Y. (2020). Microbial interaction network estimation via bias-corrected graphical lasso. *Stat Biosci.* 13, 329–350. doi: 10.1007/s12561-020-09279-y

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PloS Comput. Biol.* 11, e1004226. doi: 10.1371/journal.pcbi.1004226

Li, B., and Solea, E. (2018). A nonparametric graphical model for functional data with application to brain networks based on fMRI. *J. Am. Stat. Assoc.* 113, 1637–1655. doi: 10.1080/01621459.2017.1356726

Madan, J. C., Koestler, D. C., Stanton, B. A., Davidson, L., Moulton, L. A., Housman, M. L., et al. (2012). Serial analysis of the gut and respiratory microbiome in cystic fibrosis in infancy: interaction between intestinal and respiratory tracts and impact of nutritional exposures. *mBio* 3, e00251–e00212. doi: 10.1128/mBio.00251-12

Meinshansen, N., and Bühlmann, P. (2006). High dimensional graphs and variable selection with lasso. *Ann. Stat* 34, 1436–1462. doi: 10.1214/009053606000000281

Mohammadi, R., and Wit, E. (2019). BDgraph: an R package for bayesian structure learning in graphical models. *J. Stat. Softw.* 89, 1–30. doi: 10.18637/jss.v089.i03

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci.* 103, 8577–8582. doi: 10.1073/pnas.0601602103

Qiao, X., Guo, S., and Gareth, J. M. (2019). Functional graphical models. *J. Am. Stat. Assoc.* 114, 525. doi: 10.1080/01621459.2017.1390466

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65. doi: 10.1038/nature08821

Qiu, H., Han, F., Liu, H., and Caffo, B. (2016). Joint estimation of multiple graphical models from high dimensional time series. *J. R Stat. Soc. Ser. B Stat. Methodol.* 78, 487–504. doi: 10.1111/rssb.12123

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinf.* 12, 77. doi: 10.1186/1471-2105-12-77

Solea, E., and Li, B. (2020). Copula gaussian graphical models for functional data. *J. Am. Stat. Assoc.* 117 (538), 781–793. doi: 10.1080/01621459.2020.1817750

Staley, J. T., and Konopka, A. E. (1985). Measurement of in *situ* activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu. Rev. Microbiol.* 39, 321–346. doi: 10.1146/annurev.mi.39.100185.001541

Tian, C., Jiang, D., Hammer, A., Sharpton, T., and Jiang, Y. (2023). Compositional graphical lasso resolves the impact of parasitic infection on gut microbial interaction networks in a zebrafish model. *J. Am. Stat. Assoc.* 118, 1500–1514. doi: 10.1080/01621459.2022.2164287

Viles, W. D., Madan, J. C., Li, H., Karagas, M. R., and Hoen, A. G. (2021). Information content of high-order association of the human gut microbiota network. *Ann. Appl. Stat.* 15, 1788–1807. doi: 10.1214/21-AOAS1449

Wang, C., and Jiang, B. (2020). An efficient ADMM algorithm for high dimensional precision matrix estimation via penalized quadratic loss. *Comput. Stat Data Anal.* 142, 106812. doi: 10.1016/j.csda.2019.106812

Wille, A., and Bühlmann, P. (2006). Low-order conditional independence graphs for inferring genetic networks. *Stat. Appl. Genet. Mol. Biol.* 5. Art. 1, 34 pp. (electronic). doi: 10.2202/1544-6115.1170

Yoon, G., Gaynanova, I., and Muller, C. L. (2019). Microbial networks in SPRING-Semiparametric rank-based correlation and partial correlation estimation for quantitative microbiome data. *Front. Genet.* 10, 516. doi: 10.3389/fgene.2019.00516

Yuan, H., He, S., and Deng, M. (2019). Compositional data network analysis via lasso penalized D-trace loss. *Bioinformatics* 35, 3404–3411. doi: 10.1093/bioinformatics/btz098

Yuan, M., and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* 94, 19–35. doi: 10.1093/biomet/asm018

Zhou, J., Deng, Y., Luo, F., He, Z., Tu, Q., and Zhi, X. (2010). Functional molecular ecological networks. *mBio* 1, e00169–e00110. doi: 10.1128/mBio.00169-10

Zhu, H., Strawn, N., and Dunson, D. B. (2016). bayesian graphical models for multivariate functional data. *J. Mach. Learn. Res.* 17, 1–7.