Identifying Microbial Interaction Networks Based on Irregularly Spaced Longitudinal 16S rRNA sequence data

Jie Zhou[a,b], Jiang Gui[b,*], Weston D. Viles[d], Haobin Chen[b], Juliette C. Madan[c], Modupe O. Coker[e], Anne G. Hoen[b,c,*]

[a] *Department of Statistics, School of Mathematics and Statistics, Xidian University, Xi'an, Shaanxi, China.*
[b] *Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Hanover, NH, U.S.A*
[c] *Department of Epidemiology, Geisel School of Medicine, Dartmouth College, Hanover, NH, U.S.A*
[d] *Department of Mathematics and Statistics, University of Southern Maine, Portland, ME, U.S.A*
[e] *School of Dental Medicine, University of Rutgers, Newark, NJ, U.S.A*

Abstract

Though being vital for human health, microbial interactions with their host and with each other are still largely obscure for researchers. To deepen the understanding, the analyses based on longitudinal data are a better choice than the cross-sectional data since the information provided by the former is usually more stable. To this end, in this paper, we first propose an EM-type algorithm to identify microbial interaction network for the irregularly spaced longitudinal measurements. Correlation functions are employed to account for the correlation across the temporal measurements for a given subject. The algorithms take advantage of the efficiency of the popular graphical lasso algorithm and can be implemented straightforwardly. Simulation studies show that the proposed algorithms can significantly outperform the conventional algorithms such as graphical lasso or neighborhood method when the correlation between measurements grows larger. In second part of the paper, based on a 16S rRNA sequence data set of gut microbiome, module-preserving permutation test is proposed to test the independence of the estimated network and the phylogeny of the microbe species. The results demonstrate evidences of strong association between the interaction network and the phylogenetic tree which indicates that the taxa closer in their genomes tend to have more/stronger interactions in their functions. The proposed algorithms can be implemented through R package lglasso at `https://github.com/jiezhou-2/lglasso`

1. Introduction

Microorganisms thrive in communities with large numbers. They interact with their host and with one another in various forms such as commensalism, synergism, competition, parasitism, and predation etc. This complex set of interactions can be depicted in a form of microbial interaction network (MIN) (Faust and Raes (2012)). Traditionally, such interactions are inferred using culture based methods which can only accommodate small number of microbial strains (Gause (1934); Harcombe (2010)). Since majority of the microbes cannot be cultivated, the estimated interactions in laboratory are of limit and may

*Corresponding author

be misleading. Underpinned by the advances in next generation sequencing (NGS) technologies, complete profile of microbiome can be measured with relatively low cost nowadays and therefore provides the opportunity for the researchers to investigate the microbial interaction in real environments. However, the complexities of these high-throughput data such as high dimensionality, zero inflation and compositional nature etc compose substantial challenges for the downstream analysis such as the identification of MIN (Faust and Raes (2012)). Currently, the primary way for the inference of MIN is the so-called pairwise method in which the co-occurrence or mutual exclusion pattern of two species are compared using a measure such as Pearson or Spearman correlation (Arumugam et al (2011); Barberan (2012); Qin et al (2010); Zhou et al (2010)). A emerging method is based on the concept of conditional independence, i.e., the conditional joint distribution of two microbes given all the other microbes in the microbiome. Conditional independence is conceptually superior to the pairwise method since it remove the effects of all the other microbes when measuring the relationship between two microbes (Viles et al (2021); Chen et al (2017); Zhou et al (2020)). In particular, if we assume that the data is from a high-dimensional normal distribution, then the inverse matrix of covariance matrix, i.e, so-called precision matrices, directly reflect the conditional independence relationship among microbes. With such an appealing interpretation, precision matrices is the ideal tools to explore the MIN if the assumption of normality is reasonable in real situations.

The estimation of high-dimensional precision matrices has attracted tons of attention in literature (Wang and Jiang, 2020; Marco et al, 2018), among which the most popular way is the L_1-penalized maximum likelihood estimation. An efficient algorithm called graphical lasso is proposed to implement the L_1-penalized maximum likelihood estimation of precision matrix (Friedman et al (2008a,b)). Graphical lasso breaks down the estimation of $p \times p$ precision matrix to p L_1-penalized regression problems, to which the well-known coordinate descent algorithm are applied. Another popular yet less efficient method is the so-called neighborhood method (Meinshansen and Buhlmann (2006)). Neighborhood method estimates the neighborhood of each node and then pieces these neighborhood together to form the final estimate of network. However, one of the prerequisite conditions for these methods is that the data have to be independent and from the same normal distribution. Though such constrain may be reasonable if the data is cross-sectional, in the clinical trials such as New Hampshire Birth Cohort Study (NHBCS), many data sets are of longitudinal nature in which multiple observations are made on the same subject. In such study, the observations such as the gut microbiome from same subject are believed to be correlated and consequently the methods mentioned above do not apply to such situations. It should be noted that there have been studies in literature to estimate the network based on the longitudinal data Chen et al (2017); Sacha et al (2018). However, as far as we know, all the existing studies have employed the time series method, e.g, vector auto-regression or state-space model, to deal with the correlation between consecutive observations within the same cluster. Such methods have to assume that the data is equally spaced and the length of the data from a given subject should be long enough. These requirements are often violated for many longitudinal data set in clinical trial.

In this paper, we consider the estimation of MIN from irregularly-spaced longitudinal data based on 16S rRNA sequence data. We assume that the relative abundance, after some transformation, are normally distributed and focus on the estimation of its precision matrix. As a preliminary result, we first proposed a so-called homogeneous model in which all the subjects are assumed to share a same auto-correlation parameter τ. The parameter τ controls the extent to which the data at different times points are correlated. For this

model, an iterative algorithm is proposed to estimate the network which integrates the classic graphical lasso algorithm so that the estimator is guaranteed to be the maximum of the penalized likelihood function. The algorithm can be implemented straightforwardly based on the existing R package such as *glasso*(Friedman et al (2008a)). Based on this preliminary result, we then propose the so-called heterogeneous model in which each subject possesses his/her own auto-correlation parameter τ_i. For this model, by taking the individual τ_i as missing data, we proposed an EM type algorithm to estimate the network and τ_i's. The proposed algorithm integrates the expectation-maximization (EM) algorithm with graphical lasso and converges to the penalized maximum likelihood estimator. It can also be implemented readily based on *glasso* package. Simulation studies demonstrate the superiority of the proposed models compared to the existing methods such as graphical lasso (Friedman et al (2008b)) and neighborhood method (Meinshansen and Buhlmann (2006)).

In second part, we use the proposed models to study a longitudinal gut microbiome data set from the New Hampshire Birth Cohort Study (NHBCS), a large ongoing prospective cohort study of women and their young children Madan et al (2016); Zhou et al (2020, 2021). With the estimated MIN in hand, we then measure the correlation between the estimated MIN and their genomes through the phylogenetic tree. To this end, the length of shortest path between two taxa on the MIN (phylogenetic tree) are used as the measure of the relatedness. Then a global permutation test procedure is proposed to detect the correlation. The results demonstrate strong correlation between these MIN and phylogenetic tree which indicates that taxa that are closer in their genomes also have more/stronger interactions. Similar discoveries have been noted in other studies (Chaffron et al (2010); Eiler et al (2012)).

The paper is organized as follows. In Section 2, we introduce the main results in this study in which Section 2.1 defines the underlying data generating process for microbial abundance, Section 2.2 introduces the inference algorithm that proposed for the homogeneous data model, and Section 2.3 describe the inference algorithm for the heterogeneous data model. Section 3 compares the performance of the proposed algorithm with the conventional methods using simulated data. In Section 4, module-preserving permutation test is proposed to validate the estimated MIN. We conclude with a brief discussion in Section 5.

2. Model and Algorithms

2.1. Data generating process

Let $\mathbf{y}_{it} = (y_{it1}, \cdots, y_{itp})^T$ denote observations of some transformed abundance data of microbiome with p taxa from subject i at time t ($1 \leq i \leq m$, $t_{i1} \leq t \leq t_{in_i}$) so that it is appropriate to assume that

$$\mathbf{y}_{it} \sim N_p(\boldsymbol{\mu}, \Sigma). \tag{1}$$

where $\boldsymbol{\mu} = E(\mathbf{y}_{it})$ and $\Sigma = Var(\mathbf{y}_{it})$ are mean vector and covariance matrix respectively. Precision matrix is defined as $\Omega = \Sigma^{-1}$. Then $n_i p$ vector $\mathbf{y}_i = (\mathbf{y}_{i1}^T, \cdots, \mathbf{y}_{in_i}^T)^T$ represents all the observations for subject i, and vector $\mathbf{y} = (\mathbf{y}_1^T, \cdots, \mathbf{y}_m^T)^T$ all the observations for m subjects with $n = \sum_{i=1}^m n_i$. As for the correlation between the observations, we assume that the observations from different subjects are independent, i.e.,

$$cov(\mathbf{y}_{i_1 t_{k_1}}, \mathbf{y}_{i_2 t_{k_2}}) = \mathbf{0}_{p \times p} \tag{2}$$

for $i_1 \neq i_2, k_1 > 0, k_2 > 0$. While for the observations from the same subject, we assume

$$cov(\mathbf{y}_{it_{k_1}}, \mathbf{y}_{it_{k_2}}) \quad = \quad DH_{ik_1k_2}D \tag{3}$$

where $D = \text{diag}(\sigma_1, \cdots, \sigma_p)$ with $\sigma_1^2, \cdots, \sigma_p^2$ the diagonal elements of Σ, while $H_{ik_1k_2}$ is the correlation matrix between $\mathbf{Y}_{it_{k_1}}$ and $\mathbf{Y}_{it_{k_2}}$ for which the following form is assumed,

$$H_{ik_1k_2} \quad = \quad \Phi_{ik_1k_2} \odot R \tag{4}$$

The symbol $\odot$ stands for the element-wise product of $p \times p$ matrices $\Phi_{ik_1k_2}$ and R. Here R is the correlation matrix of covariance matrix Σ, while matrix $\Phi_{ik_1k_2} = (\phi_{ik_1k_2})_{p\times p}$ defines the dampening rates at which the absolute values of correlations decrease as time goes from t_{k_1} to t_{k_2}. For example, $(\Phi_{ik_1k_2})_{12}$ reflects the dampening rate of the absolute value of correlation $cor(Y_{it_{k_1}1}, Y_{it_{k_1}2})$ to that of correlation $cor(Y_{it_{k_1}}, Y_{it_{k_2}2})$. Theoretically dampening rates can vary from component to component, and also depend on the time points considered as long as the resulting matrix $H_{ik_1k_2}$ is positive definite. However, in this paper we will assume all the components share the same dampening rate and their dependence on time points (t_{k_1}, t_{k_2}) only through the distance between the time points, $|t_{k_1} - t_{k_2}|$. For example, the following form

$$\phi_{ik_1k_2} = \exp(-\tau_i(t_{k_1} - t_{k_2})^2) \tag{5}$$

is called Gaussian correlation function, while

$$\phi_{ik_1k_2} = \exp(-\tau_i|t_{k_1} - t_{k_2}|) \tag{6}$$

is named as exponential correlation function. Here τ_i's in (5) and (6), which are referred to as correlation parameters, are the dampening rates that shared by all the components of $\mathbf{y_{it}}, (i = 1, \cdots, n)$. One of the advantages of this model is that τ_i's reflect the subject level information which therefore can serve as a possible biomarker for some diseases in clinical trial. The choice of the form for correlation function usually has minor impact on the final conclusion. In numerical studies in Sections 3 and 4, we always employ the exponential correlation function in our model and assume that the observations have been centered so that $\boldsymbol{\mu} = \mathbf{0}$. Let Σ_i denote the covariance matrix of the observation vector $\mathbf{y}_i$ for subject i, with the assumptions (1)$\sim$(4) in hand, the density function of $\mathbf{y}$ is given by

$$f(\mathbf{y}|\Sigma_i, \boldsymbol{\tau}) = \prod_{i=1}^{n} f_i(\mathbf{y}_i|\Sigma_i, \boldsymbol{\tau}) \tag{7}$$

where $f_i(\mathbf{y}_i|\Sigma_i, \tau_i) = (2\pi)^{-n_ip/2}|\Sigma_i|^{-1/2}\exp(-\mathbf{y}_i^T\Sigma_i^{-1}\mathbf{y}_i/2)$ with

$$\Sigma_i = \begin{pmatrix} \Sigma & e^{-\tau_i|t_1-t_2|}\Sigma & \cdots & e^{-\tau_i|t_1-t_{n_1}|}\Sigma \\ e^{-\tau_i|t_2-t_1|}\Sigma & \Sigma & \cdots & e^{-\tau_i|t_2-t_{n_2}|}\Sigma \\ \vdots & \vdots & \vdots & \vdots \\ e^{-\tau_i|t_{n_i}-t_1|}\Sigma & e^{-\tau_i|t_{n_i}-t_2|}\Sigma & \cdots & \Sigma \end{pmatrix} = \Phi_i \otimes \Sigma$$

and

$$\Phi_i = (\phi_{ijk})_{n_i\times n_i} = \begin{pmatrix} 1 & e^{-\tau_i|t_1-t_2|} & \cdots & e^{-\tau_i|t_1-t_{n_1}|} \\ e^{-\tau_i|t_2-t_1|} & 1 & \cdots & e^{-\tau_i|t_2-t_{n_2}|} \\ \vdots & \vdots & \vdots & \vdots \\ e^{-\tau_i|t_{n_i}-t_1|} & e^{-\tau_i|t_{n_i}-t_2|} & \cdots & 1 \end{pmatrix}$$

Since $\Sigma_i^{-1} = \Phi_i^{-1} \otimes \Sigma^{-1}$ and $|\Sigma_i| = |\Phi_i|^p |\Sigma|^{n_i}$, the computation of the inverse matrix of $n_i p \times n_i p$ matrix Σ_i and its determination in density function (7) can be broken down to that of matrix Φ_i and Σ which can greatly reduce the computation of likelihood function (7). Furthermore, such form of likelihood function enables us to employ the efficient graphical lasso to compute the penalized maximum likelihood estimator of underlying network, i.e., precision matrix $\Omega = \Sigma^{-1}$.

In the following two sections, we consider the selection the underlying network Ω. Given the fact that the number of unknown parameters are usually much higher than sample size in the context of gut microbiome, the sparsity is typically assumed for the structure of Ω. Under the sparsity assumption, if the observations are independent, the graphical lasso algorithm (Friedman et al (2008a)) are the most popular algorithm which yields the penalized likelihood estimation of Ω. Another popular method is the neighborhood method (Meinshansen and Buhlmann (2006)) which, though easier to be implemented, are not the most efficient since it is not based on likelihood function. For correlated data, the existing studies focus on the inference of network defined by the state transition matrix between consecutive time points in time series models (Chen et al (2017); Epskamp et al (2018)). Consequently, the times series in these studies have to be long enough and equally spaced. Longitudinal data in clinical trial such as NHBCS usually have shorter length and are irregularly spaced which does not satisfy these requirements. With these observations in mind, we propose an algorithm to identify Ω in model (7). To ease the exposition, an algorithm for homogeneous model are considered first based on which the algorithm for the general case in (7) are proposed. Both algorithms yield the penalized likelihood estimates of Ω.

2.2. Network with homogeneous correlation parameter

In this section, we consider identifying Ω under the assumption $\tau_1 = \cdots = \tau_m = \tau$ which says that correlations of microbiota for each subjects in the cohort dampens at the same rate. From the density function (7), the log likelihood function for $\mathbf{y} = (\mathbf{y}_1^T, \cdots, \mathbf{y}_m^T)^T$ is given by

$$l_n(\Sigma, \tau|\mathbf{y}) = -\frac{1}{2}\sum_{i=1}^{m}\left(p\log(|\Phi_i|) + n_i\log(|\Sigma|) + \mathbf{y}_i^T(\Phi_i \otimes \Sigma)^{-1}\mathbf{y}_i\right) \tag{8}$$

up to a constant. Here we have used the fact that $\Sigma_i = \Phi_i \otimes \Sigma$ and $|\Sigma_i| = |\Phi_i|^p|\Sigma|^{n_i}$. Note that the last term in log likelihood function (8) can be written as in the following form,

$$\begin{aligned}
\mathbf{y}_i^T(\Phi_i \otimes \Sigma)^{-1}\mathbf{y}_i &= \mathbf{y}_i^T(\Phi_i^{-1} \otimes \Sigma^{-1})\mathbf{y}_i = \sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\phi_{ijk}^{-}\mathbf{y}_{ij}^T\Omega\mathbf{y}_{ik} \\
&= \sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\mathrm{tr}(\phi_{ijk}^{-}\mathbf{y}_{ik}\mathbf{y}_{ij}^T\Omega) = \mathrm{tr}\left(\left(\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\phi_{ijk}^{-}\mathbf{y}_{ik}\mathbf{y}_{ij}^T\right)\Omega\right) \\
&\triangleq \mathrm{tr}S_i\Omega, \qquad (9)
\end{aligned}$$

where $S_i(\tau) = \sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\phi_{ijk}^{-}\mathbf{y}_{ik}\mathbf{y}_{ij}^T$ and $\Phi_i^{-1} = (\phi_{ijk}^{-})_{n_i \times n_i}$. Substituting (9) into (8), we have

$$l_n(\Omega, \tau|\mathbf{y}) = -\frac{1}{2}\left\{\sum_{i=1}^{m}p\log(|\Phi_i|) - n\log(|\Omega|) + n\mathrm{tr}(\bar{S}(\tau)\Omega)\right\}, \tag{10}$$

where $n = \sum_{i=1}^m n_i, \bar{S}(\tau) = \frac{1}{n}\sum_{i=1}^m S_i(\tau)$. Here we used $\bar{S}(\tau)$ to emphasize that matrix $\bar{S}$ is a function of unknown parameter τ. Consequently, a sparse network can be achieved by minimizing the following L_1-penalized negative likelihood function,

$$\min_{\Omega,\tau}\{-2l_n(\Omega,\tau|\mathbf{y}) + n\lambda|\Omega|_1\}, \tag{11}$$

for given tuning parameter $\lambda > 0$. The maximization problem (11) can be carried out through an iterative way. First note for given τ, the solution of Ω can be obtained through the following minimization,

$$\min_{\Omega}\{-\log|\Omega| + \mathrm{tr}(\bar{S}(\tau)\Omega) + \lambda|\Omega|_1\}, \tag{12}$$

which happens to has the same form as the conventional graphical lasso for independent data when the empirical covariance matrix is given by $\bar{S}(\tau)$. Consequently, graphical lasso algorithm can be used to compute the sparse estimate of Ω in (12). On the other hand, given Ω, the maximization of (11) with respect to τ does not involve any l_1 penalty term and consequently can be carried out through the maximization of likelihood function (10) with respect to τ. The conventional Newton algorithm can be used in this step. This process is proceeded until the convergence is achieved. The algorithm is summarized in Algorithm 1.

Algorithm 1 Identify the network for homogeneous model

1: **procedure**
2: Given initial value τ_0 and Ω_0, tuning parameter λ and error tolerance $e > 0$:
3: With $\tau = \tau_0$, solve optimization problem (12) with respect to Ω using graphical lasso and let $\hat{\Omega}$ be the resulting estimate of Ω.
4: With $\Omega = \hat{\Omega}$, solve optimization problem (10) with respect to τ. Let $\hat{\tau}$ be the resulting estimate of τ.
5: **if** $|\tau_0 - \hat{\tau}| < e$ and $|(\Omega_0 - \hat{\Omega})ij| < e$ for $1 \le i \le j \le p$ **then**
6: Stop and output $(\hat{\tau}, \hat{\Omega})$.
7: **else**
8: Let $\tau_0 = \hat{\tau}, \Omega_0 = \hat{\Omega}$, return to line 3.

2.3. Network with heterogeneous correlation parameter

In previous section, we have assumed that a single correlation parameter τ applies to all the individuals. In clinical trial, this parameter may potentially be affected by some covariates such as gender, diet etc. and in turn impact the health outcome and so τ's themselves are also of interest. Here we consider the network identification without assuming $\tau_1 = \cdots = \tau_m$. Specifically, we assume the parameters τ_i's are independent random variables with identical exponential distribution, $\tau_i \sim \exp(\alpha)$ with parameter α. Consequently, the joint density function for $\{\mathbf{y}_i, \tau_i\}_{i=1}^m$ can be shown to be

$$\prod_{i=1}^m f_i(\mathbf{y}_i|\Sigma,\tau_i)\alpha\exp(-\alpha\tau_i) \tag{13}$$

from which the likelihood function is given by

$$L_n(\Omega,\alpha|\mathbf{y}) = \iiint_{\tau_1,\cdots,\tau_m}\prod_{i=1}^m f_i(\mathbf{y}_i|\Sigma,\tau_i)\alpha_i\exp(-\alpha_i\tau_i)d\tau_1\ldots\tau_m \tag{14}$$

The sparse estimates for the network can then be obtained through minimizing the following L_1-penalized negative loglikelihood function,

$$\min_{\Omega,\alpha}\{-2l_n(\Omega,\alpha)+n\lambda|\Omega|_1\}, \tag{15}$$

where $l_n(\Omega,\alpha)=\log(L_n(\Omega,\alpha|\mathbf{y}))$. Since the explicit form of $l_n(\Omega,\alpha)$ is unavailable, Expectation-Maximum (EM) algorithm are proposed here to find the penalized maximum likelihood estimate of Ω and α (Dempster et al (1977)). Since we are considering the negative likelihood function in (15), so the maximization in EM algorithm will be replaced by minimization here. The correlation parameters $\boldsymbol{\tau}=(\tau_1,\cdots,\tau_m)$ will be taken as the so-called missing data. Recall in the first step of EM algorithm, the conditional distribution of missing data $\boldsymbol{\tau}$ given $\mathbf{y},\Sigma=\Sigma_0,\alpha=\alpha_0$ has to be derived which from (7) and (13) can be shown to be

$$\begin{aligned} g(\boldsymbol{\tau}|\mathbf{y},\Sigma_0,\alpha_0) &= \prod_{i=1}^m g(\tau_i|\mathbf{y}_i,\Sigma_0,\alpha_0) \\ &\propto \prod_{i=1}^m |\Phi_i|^{-p/2}\exp(-\mathbf{y}_i^T(\Phi_i^{-1}\otimes\Sigma_0^{-1})\mathbf{y}_i/2)\exp(-\alpha_0\tau_i). \end{aligned} \tag{16}$$

With (16) in hand, the expectation of (15) with respect to (16) is

$$\begin{aligned} Q(\Omega,\alpha|\Omega_0,\alpha_0) &= \sum_{i=1}^m pE_g\{\log(|\Phi_i|)\}-n\log(|\Omega|)+n\mathrm{tr}\left(S^{(0)}\Omega\right) \\ &\quad +m\log(\alpha)-\alpha\sum_{i=1}^m E_g\tau_i+\lambda|\Omega|_1 \end{aligned} \tag{17}$$

where $S^{(0)}=E_g\bar{S}(\boldsymbol{\tau})$. In the second step of EM algorithm, the minimum of the Q function (17) has to be computed. This again is implemented through an iterative manner. First, for fixed Ω, it is straightforward to show that the minimum of Q function with respect to α is attained at

$$\hat{\alpha}=\frac{1}{\frac{1}{m}\sum_{i=1}^m E_g\tau_i}, \tag{18}$$

i.e., the reciprocal of the sample mean of the conditional expectation of τ_i. Then, for given α, the minimization of (17) is equivalent to,

$$\min_{\Omega}\{-\log|\Omega|+\mathrm{tr}\left(S^{(0)}\Omega\right)+\lambda|\Omega|_1\} \tag{19}$$

which can be solved directly through graphical lasso algorithm. The difficult part of this algorithm is to find the expectation $E_g\bar{S}(\boldsymbol{\tau})$, which may not admit an explicit form given the fact that $\bar{S}(\boldsymbol{\tau})$ is a nonlinear function of $\boldsymbol{\tau}$ and the complex form of density function $g(\boldsymbol{\tau})$; therefore this expectation will be computed through Monte Carlo method. The details of the algorithm are summarized in Algorithm 2.

For a given sequence of tuning parameters λ's, a solution path can thus be generated. In order to select the optimal network Ω from this model pool, a variety of model selection criteria can be used, e.g, Akaike information criterion (AIC), Bayesian information criterion (BIC) or cross validation (CV) etc. In the numerical studies in the next two sections, we use the extended BIC (EBIC) proposed in Foygel and Drton (2010); Zhou et al (2021) to select the appropriate tuning parameter.

Algorithm 2 Identify the network for heterogeneous model

1: **procedure**
2: Given initial value α_0 and Ω_0, tuning parameter λ and error tolerance $e > 0$
3: Sample $\tau_i (i = 1, \cdots, m)$ from the distribution (16) with $\Omega = \Omega_0, \alpha = \alpha_0$. Let the samples denote by $\boldsymbol{\tau}_i = (\tau_{i1}, \cdots, \tau_{in})(1 \leq i \leq m)$.
4: Estimate $E_g \tau_i, E_g S_i(\tau_i)(i = 1, \cdots, m)$ using the corresponding empirical quantities $\bar{\tau}_i, \bar{S}_i(\tau_i)$.
5: Estimate α by(18) and Ω by $\hat{\Omega}$, the solution of (19) with $E_g S_i(\tau_i)$ replaced by $\bar{S}_i(\tau_i)$.
6: **if** $|\alpha_0 - \hat{\alpha}| < e$ and $|(\Omega_0 - \hat{\Omega})_{ij}|_1 < e$ **then**
7: Stop and output $(\hat{\alpha}, \hat{\Omega}, \hat{\tau}_i, i = 1, \cdots, m)$.
8: **else**
9: $\alpha_0 = \hat{\alpha}$ and $\Omega_0 = \hat{\Omega}$, return to line 3.

3. Simulation

In this section, we compare the proposed algorithm with conventional network selection methods, such as graphical lasso (Friedman et al (2008a, 2019)), neighborhood algorithm (Meinshansen and Buhlmann (2006)) under different parameter settings, form which it is demonstrated that the proposed temporal grpahical lasso algorithm can outperform these existing algorithms when the correlations between the consecutive observations are not negligible. Motivated by the indices true positive rate (TPR) and false positive rate (FPR), the following two indices are used to measure the difference of two given network G_1 and G_2 which share the same nodes,

$$\mathrm{R}_1(G_1, G_2) = \frac{|G_1 \cap G_2|}{|G_2|}, \quad \mathrm{R}_2(G_1, G_2) = \frac{|G_1 - G_2|}{|\bar{G}_2|}, \tag{20}$$

where $|G_2|$ represents the number of edges in G_2, $|\bar{G}_2|$ the number of null edges in G_2, $|G_1 \cap G_2|$ the number of shared edges by G_1 and G_2 and $|G_1 - G_2|$ the number of edges that appear in G_1 while not in G_2. It is straightfoward to show that in case of $R_1 = 1$ and $R_2 = 0$, these two network G_1 and G_2 just share the same edge set, i.e., $G_1 = G_2$. Paricularly, when G_2 is the underlying true network and G_1 represents the estimated network, then R_1 and R_2 are just the conventional TPR and FPR respectively.

First we show that the network estiamted by temporal graphical lasso will coincide with the network estimated by graphical lasso when the correlation between consecutive observations tends to zero. Specifically, we consider a network with $p = 100$ nodes that is shared by $m = 20$ subjects. The precision matrix corresponding to this network is generated through the R package *BDgraph* with edge density equal to 0.2. For each subject, observations along four time points are generated. For subject i, the space between two consecutive time points t_{ij} and $t_{i(j+1)}$ is generated by $\max\{\Delta_{ij}, 0.5\}$ where Δ_{ij}'s follow a Poisson distribution with mean value 1. Then the data for subject i is generated with correlation parameter $\tau_0 = 0.5$. For this dataset, the temporal graphical lasso and graphical lasso are employed respectively to estimate the underlying network. Let G_1 and G_2 denote the resulted network from these two algorithms with a given tuning parameter λ.

Figure 1 plots the indices R_1 and R_2 as functions of correlatoin parameter τ based on 50 replicates with tuning parameter $\lambda = 0.1$. It can be seen that for large τ, the differece between G_1 and G_2 is small. In particular, for $\tau > 3$, there is little difference between t G_1 and G_2. However, for small τ, these two algorithm can generate radically different network.

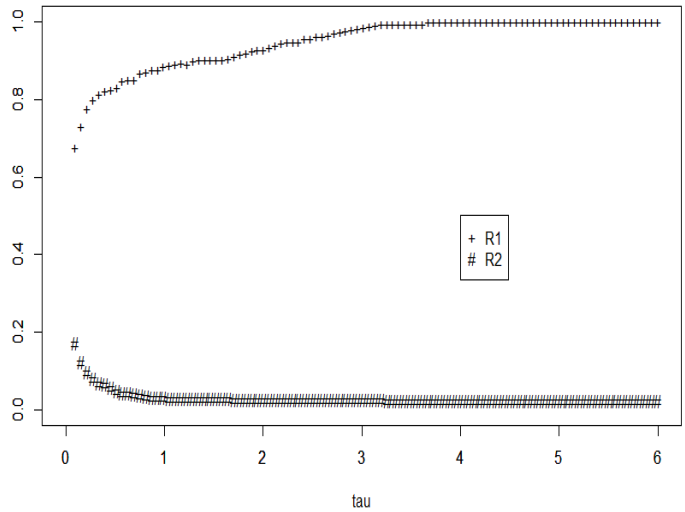

Figure 1: The difference of network between temporal graphical lasso and graphical lasso as function of τ. Here tuning parameter is set to be $\lambda = 0.1$

This phenomenon is reasonable since from equation (7), as τ becomes large, the data at different time points will tend to be independent. For independent data, our algorithm will yield the same results as graphical lasso.

In the following, we compare the performance of temporal graphical lasso, graphical lasso and neighborhood method for homogeneous and heterogeneous model respectively with EBIC as the model selection criterion (Foygel and Drton (2010); Quang et al (2021); Zhou et al (2021)). For homogeneous model, the numbers of subjects considered are $m = 15, 30$ respectively. In each case, the numbers of time points are set to be 7 and 14 respectively. Three choices of correlation parameter are $\tau = 0.1, 0.4, 0.8$. For each combination of these parameters, the data is generated by the same procedure as above. The results for the three algorithms as mentioned above are listed in Table 1. From Table 1, it can be seen that for marjority of the cases considered, the proposed temporal graphical lasso has the smallest FPRs and highest TPRs. Another observation is that neighborhood method seems to outperform graphical lasso and more robust with regard to the correlation between the data. This may be attributed to the fact that, contrary to graphical lasso, neighborhood method does not rely on the likelihood function of the data directly and therefore less impacted by the correlation between the data. As for heterogeneous model, the number of subjects considered is set to be 20. Three numbers of time points are set to be $n = 10, 20, 30$. The parameter α in exponential distribution (13) are set to be $\alpha = 1.25, 2.5, 10$, which corresponds to $\tau = 0.8, 0.4, 0.1$ respectively in Table 1. For each of these three α, 20 random samples are generated form the corresponding exponential distribution which will be taken as the correlation parameters for the 20 subjects. The spaces between consecutive time points are generated using the same procedure as above. From the heterogeneous temporal graphical lasso, we obtain two types of results, one is the estimated network and the others

Table 1: Comparison of three different network selection algorithms, temporal graphical lasso, graphical lasso and neighborhood method assuming homogeneous correlation parameters.

	$\tau = 0.1$		$\tau = 0.4$		$\tau = 0.8$	
$m = 15, n_i = 7$	TPR	FPR	TPR	FPR	TPR	FPR
glasso	0.622	0.207	0.672	0.150	0.715	0.123
nh	0.628	0.132	0.702	0.122	0.745	0.115
temglasso	0.752	0.075	0.737	0.075	0.723	0.068
$m = 15, n_i = 14$	TPR	FPR	TPR	FPR	TPR	FPR
glasso	0.744	0.258	0.785	0.164	0.783	0.113
nh	0.678	0.125	0.729	0.082	0.754	0.062
temglasso	0.802	0.066	0.796	0.065	0.792	0.061
$m = 30, n_i = 7$	TPR	FPR	TPR	FPR	TPR	FPR
glasso	0.758	0.257	0.795	0.160	0.807	0.119
nh	0.738	0.122	0.776	0.084	0.731	0.056
temglasso	0.824	0.071	0.814	0.069	0.801	0.067
$m = 30, n_i = 14$	TPR	FPR	TPR	FPR	TPR	FPR
glasso	0.776	0.206	0.798	0.087	0.772	0.045
nh	0.716	0.099	0.721	0.041	0.747	0.022
temglasso	0.875	0.040	0.869	0.041	0.865	0.041

are the predictions of the 20 τ_i's. Table 2 lists the results about the estimated network from temporal graphical lasso, graphical lasso and neighborhood method respectively. It can be seen that heterogeneous temporal graphical lasso outperforms the other two algorithms in the majority of cases considered, with highest TPR and lowest FPR. As for the prediction of τ_i, we consider two indices to evaluate the performance of predictors. One is the average relative error (ARE) defined as ARE $= \frac{1}{m}\sum_{i=1}^{m}\frac{|\tau_i - \hat{\tau}_i|}{\tau_i}$. The other is the correlation between the true τ's and the predicted value $\hat{\tau}$'s. These two indices are listed in Table 3, from which we can see that, for a given α, ARE decreases while correlation increases as sample size increases. In particular, the estimates of correlation parameter is strongly correlated with the true values of τ_i and we have observed that such correlation is rather robust to the choice of tuning parameter of λ; consequently, as an subject level index, the robust predictor $\hat{\tau}_i$ may serve as a biomarker for some clinical outcome.

4. Gut Microbial Interaction Network and Phylogenetic Tree

In this section, a longitudinal data set from the New Hampshire Birth Cohort Study (NHBCS) is investigated using the proposed heterogeneous model in Section 2. NHBCS is an ongoing prospective cohort study of women and their young children. For this data set, thirty-eight infants are involved in which the number of time points for the infants ranges from 6 to 51 months. Each observation consists of 16,383 amplicon sequence variants (ASV) for 16S rRNA gene. These sequences are then collapsed to the genus level taxa using R package *DADA2* (Callahan et al (2016)) with the reference fasta file *rdp_train_set_18.fa.gz*. The sequences that have no genus level information are dropped. Then all the taxa with the proportion of nonzero observations less than 10% are combined together which is referred

Table 2: Comparison of three different network selection algorithms, temporal graphical lasso, graphical lasso and neighborhood method assuming heterogeneous correlation parameters.

	$\alpha = 1.25$		$\alpha = 2.50$		$\alpha = 10$	
$m = 20, n = 10$	TPR	FPR	TPR	FPR	TPR	FPR
glasso	0.667	0.133	0.678	0.194	0.632	0.242
nh	0.737	0.128	0.725	0.167	0.671	0.175
temglasso	0.779	0.128	0.813	0.167	0.837	0.169
$m = 20, n = 20$	TPR	FPR	TPR	FPR	TPR	FPR
glasso	0.695	0.114	0.726	0.149	0.633	0.196
nh	0.790	0.101	0.788	0.135	0.745	0.208
temglasso	0.756	0.052	0.810	0.072	0.851	0.118
$m = 20, n = 30$	TPR	FPR	TPR	FPR	TPR	FPR
glasso	0.744	0.116	0.714	0.177	0.708	0.206
nh	0.821	0.084	0.753	0.101	0.733	0.182
temglasso	0.801	0.038	0.815	0.053	0.873	0.073

Table 3: Evaluation of the predicts of correlation parameter τ_i's in heterogeneous model. Here RE stands for the relative error, and Cor stands for the correlation between the true τ_i's and $\hat{\tau}_i$'s.

	$\alpha = 1.25$		$\alpha = 2.5$		$\alpha = 10$	
	ARE	Cor	ARE	Cor	ARE	Cor
$m = 20, n = 10$	0.235	0.980	0.253	0.982	0.453	0.965
$m = 20, n = 20$	0.133	0.987	0.135	0.992	0.232	0.987
$m = 20, n = 30$	0.105	0.992	0.109	0.995	0.164	0.993

to as composite taxon. This leaves us 83 taxa in total. The observations of zeros for each of these 83 taxa are replaced by the minimum abundance of that taxon divided by 10. The log-ratio transformation is then carried out for the relative abundance in which the composite taxon is taken as the reference. Application of the heterogeneous model in Section 2.3 to the transformed data generates the final estimated network which is displayed in Figure 2. The corresponding predicted correlation parameters τ's are shown in Figure 3. Based on the modularity maximization algorithm (Newman (2006); Vincent et al (2008)), eight communities are identified in the estimated network which are listed in Table 4. This step is accomplished through R package *igraph* Csardi (2006).

In order to demonstrate the estimated network does reflect the true structure of the underlying network, we investigate the relationship between this estimated network and phylogenetic tree of the 82 taxa (the composite one is excluded here). The pholygenetic tree constructed from the same data set is presented in Figure 4. Our underlying theory is that microbial taxa with closer evolution histories also have more/stronger interactions in human body. To find the supportive evidence from the data, the null hypothesis is set to be the estimated network has zero correlation with the phylogenetic tree. To test this hypothesis, we compute the distance between two taxa in both the estimated network and phylogenetic tree. Here the distance between taxa A and B is defined as the length of the shortest path from A to B. Let d_1 and d_2 be respectively the distance variables from the estimated network and phylogenetic tree, then the correlation between d_1 and d_2 can be used to measure the relatedness between the network and phylogenetic tree. This correlation turns out to be $r_0 = 0.228$. Five thousand bootstrap samples yields the 95% confidence interval of r_0 to be $[0.196, 0.261]$, see Figure 5.

In order to understand the significance of r_0, we use permutation to construct the reference distribution. Out of multiple ways to carry out the permutation test, we recommend a module-preserving permutation test. Specifically, we keep the structure of the estimated network unchanged, and permute the order of the 82 taxa $m = 5000$ times. For the ith permuted network, a distance $d_1^{(i)}$ is computed and consequently we have $r^{(i)} = cor(d_1^{(i)}, d_2), i = 1, \cdots, 5000$. Note that the studies in Diciccio and Romano (2017) show that it is not appropriate to use these permuted correlation $r^{(i)}$ directly to determine the significance of r_0 since it can lead to misleading p-value and large type 3 error rate. Here we adopt the proposal in Diciccio and Romano (2017) and use the following more robust studentized correlation,

$$\tilde{r} = \frac{\sqrt{l}r}{\sqrt{\frac{\hat{\mu}_{2,2}}{\hat{\mu}_{2,0}\hat{\mu}_{0,2}}}} \tag{21}$$

where $\hat{\mu}_{r,s} = \frac{1}{l}\sum_{i=1}^{l}(d_1^{(i)} - \bar{d}_1)^r(d_2^{(i)} - \bar{d}_2)^s$, l is the length of vector $d_1(d_2)$ and $\bar{d}_1(\bar{d}_2)$ are the mean value of $d_1(d_2)$. Based on (21), the p-value of $\tilde{r}_0$ turns out to be zero, which means that the correlation between the estimated network is statistically significant, i.e., the data supports the theory that microbial taxa with closer evolution histories also tend to have more/stronger interactions, see Figure 6. It should be noted similar ideas have been proposed in other studies. For example, Chaffron et al (2010) found that genomes from coexisting microbes 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 was often dependent on taxonomic relatedness. Our studies validate this theory using proposed algorithm and ASV data set.

Table 4: Eight communities selected by maximizing the modularity function of the estimated microbial interaction network in Figure 2.

C1	"Anaerostipes" "Agathobacter" "Faecalibacterium" "Dorea" "Faecalibacillus" "Anaerobutyricum" "Roseburia" "Ruminococcus2" "Dialister" "Lachnospira",
C2	"Enterobacter" "Sellimonas" "Citrobacter" "Clostridium.XVIII" "Leclercia"
C3	"Fusicatenibacter" "Collinsella" "Faecalimonas" "Mediterraneibacter" "Peptostreptococcus" "Butyricicoccus" "Corynebacterium"
C4	"Bifidobacterium" "Blautia" "Akkermansia" "Intestinibacter" "Romboutsia" "Terrisporobacter" "Turicibacter" "Eisenbergiella" "Clostridium.IV"
C5	"Clostridium.sensu.stricto" "Klebsiella" "Enterococcus" "Lactococcus" "Staphylococcus" "Prevotella" "Sutterella" "Haemophilus" "Alistipes" "Sarcina" "Monoglobus" "Gemmiger" "Dysosmobacter" "Leuconostoc" "Bilophila" "Faecalicatena" "Neisseria"
C6	"Anaerococcus" "Peptoniphilus" "Negativicoccus" "Finegoldia"
C7	"EscherichiaShigella" "Schaalia" "Granulicatella" "Gemella" "Pseudescherichia" "Rothia"
C8	"Erysipelatoclostridium" "Bacteroides" "Tyzzerella" "Megasphaera" "Enterocloster" "Hungatella" "Flavonifractor" "Clostridioides" "Clostridium.XlVa" "Fusobacterium"

Note the results above correspond to the single tuning parameter selected by EBIC. In order to compare the estimated network with networks generated from other choices of tuning parameter, for each of 40 tuning parameters ranging from $\lambda = 0.1$ to $\lambda = 60$, we carried out the same heterogeneous model and the permutation test. The estimated and permuted correlations are displayed in Figure 7. The boxplots in Figure 7, from left to right, correspond to the tuning parameters from $\lambda = 0.1$ to $\lambda = 60$ respectively. For each boxplot, the dot linked by the line is the correlation between the estimated MIN and phylogenetic tree while others correspond to the correlations generated from the permutation. The most prominent feature of Figure 7 is that the curve increases first and then decreases with the maximum around the one selected by EBIC. Note the first four and last two correlations are not significant while all the other correlations are significant under the significance level 0.05. This is reasonable since too small (large) tuning parameters lead to too dense (sparse) network which in turn yield the distances computed from the network are almost the same value (1 or ∞). Figure 7 therefore validates the effectiveness of EBIC.

5. Conclusion

We proposed algorithms to estimate the microbial interaction network based on the irregularly spaced longitudinal 16S rRNA sequence data. These algorithms show their advantages over the traditional methods where the correlations are just omitted or handled using classic time series model. For the estimated network, a module-preserving permutation test is designed to test its effectiveness, which discovered the relationship between MIN and phylogenetic tree and validated the proposed algorithms. It should be noted that we have assumed that the microbes of interest in one subject share the same time dampening rate. Though it seems to work pretty well for the data sets considered, there may be occasions in which different taxa need different time dampening rates. It is interesting to find the ways expanding the proposed algorithms to such situations in the future work.

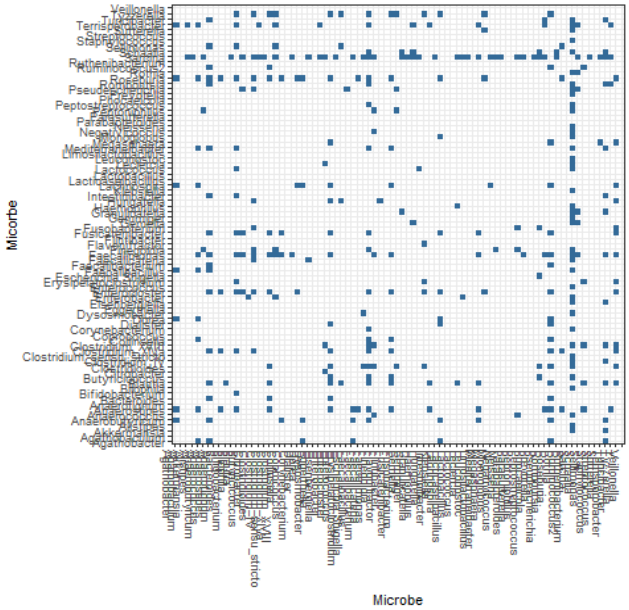

Figure 2: Microbial interaction network generated by heterogeneous temporal graphcial lasso based on the longitudinal data set in Section 4.

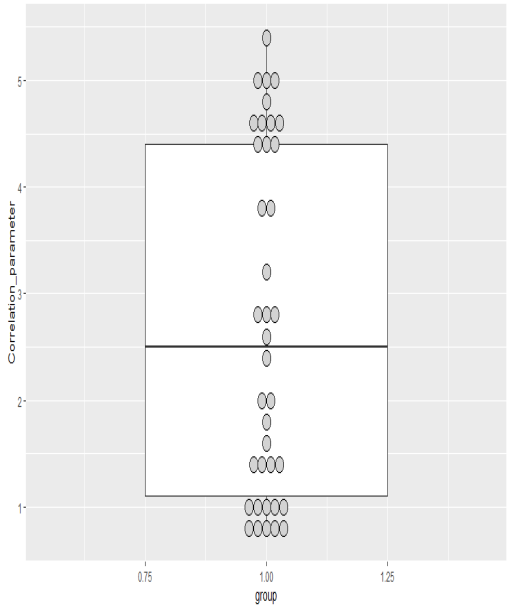

Figure 3: Dotplot for the 38 individual correlation parameter τ_i's for 38 subjects involved in the longitudinal data set in Section 4.

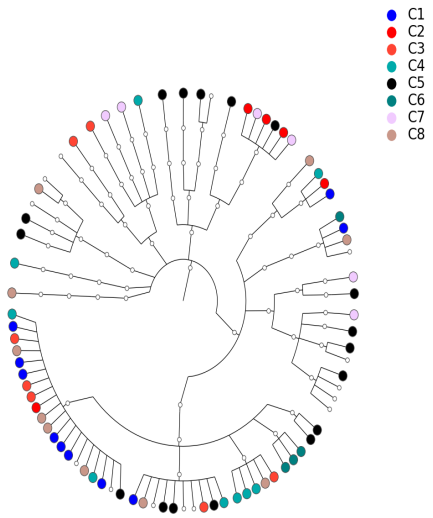

Figure 4: Phylogenetic tree for the 82 taxa involved in the longitudinal data set in Section 4. Dots with different color correspond to different communities in Table 4.

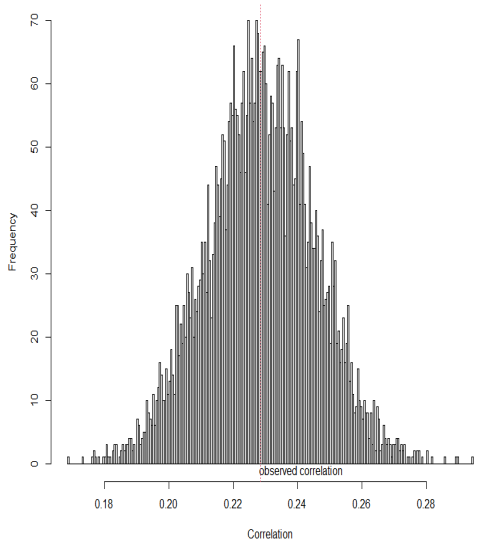

Figure 5: Bootstrapped correlations. The correlations are between taxa distances in the interaction network in Figure 2 and phylogenetic tree in Figure 4.

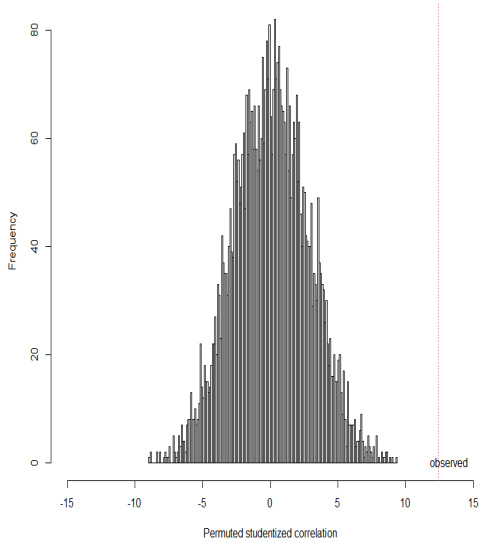

Figure 6: Permutation test for the correlation between taxa distances in the interaction network and taxa distances in the phylogenetic tree.

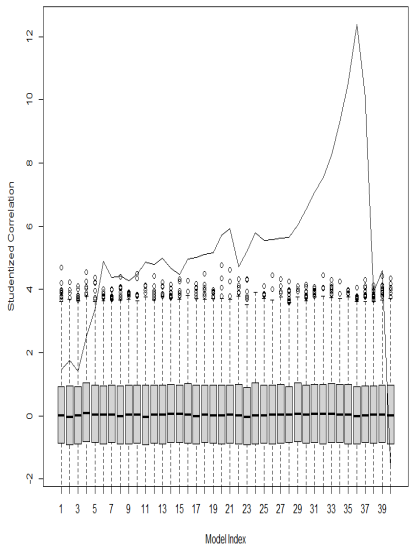

Figure 7: Forty boxplots of correlation between estimated network and phylogenetic tree, which correspond to, from the left to right, the tuning parameter λ increasing from 0.1 to 60. For a given boxplot, the dots linked are the estimated correlations from the estimated network; others are correlations computed from a 5000 permutations of taxa.

Acknowledgments

This work is supported in part by US National Institutes of Health grants (R01LM012723; P20ES018175; P01ES022832;), US Environmental Protection Agency grant (RD83459901), the Children's Center grant, the Superfund grant and COBRE.

References

Arumugam, M. et al. Enterotypes of the human gut microbiome. Nature 473, 174–180 (2011).

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

Callahan, B., McMurdie, P., Rosen, M. et al. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 13, 581–583 (2016). https://doi.org/10.1038/nmeth.3869

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

Csardi G, Nepusz T (2006). "The igraph software package for complex network research." InterJournal, Complex Systems, 1695. https://igraph.org.

Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data Via the EM Algorithm. J R Stat Soc Ser B Methodol. 1977;39(1):1-22. doi:10.1111/j.2517-6161.1977.tb01600.x

Eiler, A., Heinrich, F., Bertilsson, S. Coherent dynamics and association networks among lake bacterioplankton taxa. ISME J. 6, 330–342 (2012)

Friedman J., Hastie T., Tibshirani R. (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistcs, 9, 432-441.

Friedman J., Hastie T., Tibshirani R. (2008) Graphical Lasso: Estimation of Gaussian Graphical Models, Version: 1.11, URL: https://CRAN.R-project.org/package=glasso.

Quang Nguyen, Jie Zhou, Anne Hoen, Jiang Gui. SBICgraph: Structural Bayesian Information Criterion for Graphical Models. Version: 1.0, URL: https://CRAN.R-project.org/package=SBICgraph .

Chen I., Yogeshwar D. Kelkar., Gu Y., Zhou J., Qiu X., Wu HL. (2017) High-dimensional linear state space models for dynamic microbial interaction networks. PLoS ONE, 15: 1-20.

Cheng Wang, Binyan Jiang, An efficient ADMM algorithm for high dimensional precision matrix estimation via penalized quadratic loss, Computational Statistics and Data Analysis, Volume 142, 2020, 106812, ISSN 0167-9473, https://doi.org/10.1016/j.csda.2019.106812.

Christine Peterson, Francesco Stingo., Marina Vannucci (2015) Bayesian Inference of Multiple Gaussian Graphical Models, Journal of the American Statistical Association, 110(509), 159–174.

Diciccio CJ, Romano JP. Robust Permutation Tests For Correlation And Regression Coefficients. J Am Stat Assoc. 2017;112(519):1211-1220. doi:10.1080/01621459.2016.1202117

Epskamp S, Waldorp LJ, Mõttus R, Borsboom D. The Gaussian Graphical Model in Cross-Sectional and Time-Series Data. Multivariate Behav Res. 2018 Jul-Aug;53(4):453-480. doi: 10.1080/00273171.2018.1454823. Epub 2018 Apr 16. PMID: 29658809.

Faust K, Raes J. Microbial interactions: from networks to models. Nature Reviews Microbiology. 2012; 10(8):538–550. https://doi.org/10.1038/nrmicro2832 PMID: 22796884

Friedman J., Hastie T., Tibshirani R. (2008) R package *glasso*, URL: *http://www-stat.stanford.edu/tibs/glasso*, Version: 1.11.

Friedman J., Hastie T., Tibshirani R. (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 9(3):432-41.

Foygel R, Drton M. Extended Bayesian Information Criteria for Gaussian Graphical Models. In: Advances in Neural Information Processing Systems. Vol 23. Curran Associates, Inc.; 2010. Accessed November 21, 2021. https://papers.nips.cc/paper/2010/hash/072b030ba126b2f4b2374f342be9ed44-Abstract.html

Gause, G. F. The Struggle for Existence (Williams and Wilkins, 1934).

Geisser, S. (1975). The predictive sample reuse method with applications. J. Amer. Statist. Assoc. 70, 320-328.

Harcombe, W. Novel cooperation experimentally evolved between species. Evolution 64, 2166–2172 (2010).

Hojsgaad S., Edwards D., Lauritzen S. (2012) Graphical Models with R. New York: Springer.

Lauritzen S L. (1996). Graphical Models. Oxford University Press.

Lauritzen, S. L. and Sheehan, N. A. (2003). Graphical Models for Genetic Analyses. Statistical Science, 18(4): 489–514. MR2059327. doi: http://dx.doi.org/10.1214/ss/1081443232.

Leday G. G. and Richardson S. (2019). Fast Bayesian inference in large Gaussian graphical models. Biometrics, 75(4): 1288-1298.

Li M, Wang B, Zhang M, Rantalainen M, Wang S, Zhou H, Zhang Y, Shen J, Pang X, Zhang M, Wei H, Chen Y, Lu H, Zuo J, Su M, Qiu Y, Jia W, Xiao C, Smith LM, Yang S, Holmes E, Tang H, Zhao G, Nicholson JK, Li L, Zhao L. (2008). Symbiotic gut microbes modulate human metabolic phenotypes. Proc Natl Acad Sci, 105(6):2117-22.

Madan JC, Hoen AG, Lundgren SN, Farzan SF, Cottingham KL, Morrison HG, Sogin ML, Li H, Moore JH, Karagas MR. (2016) Association of Cesarean Delivery and Formula Supplementation With the Intestinal Microbiome of 6-Week-Old Infants. JAMA Pediatr, 170(3):212-9.

Marino S, Baxter NT, Huffnagle GB, Petrosino JF, Schloss PD. (2014) Mathematical modeling of primary succession of murine intestinal microbiota. Proceedings of the National Academy of Sciences, 111 (1): 439–444.

Marco Avella-Medina, Heather S Battey, Jianqing Fan, Quefeng Li, Robust estimation of high-dimensional covariance and precision matrices, Biometrika, Volume 105, Issue 2, June 2018, Pages 271–284, https://doi.org/10.1093/biomet/asy011

Meinshansen N., P Buhlmann. (2006). High dimensional graphs and variable selection with lasso. The annals of statistics, 34(3), 1436–1462.

Newman MEJ. Modularity and community structure in networks. Proc Natl Acad Sci. 2006;103(23):8577-8582. doi:10.1073/pnas.0601602103

Qin, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464, 59–65 (2010).

Sacha Epskamp, Lourens J. Waldorp, Rene Mottus Denny Borsboom (2018) The Gaussian Graphical Model in Cross-Sectional and Time-Series Data, Multivariate Behavioral Research, 53:4, 453-480, DOI: 10.1080/00273171.2018.1454823

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre: Fast unfolding of communities in large networks. J. Stat. Mech. (2008) P10008

Verbeke G, Fieuws S, Molenberghs G, Davidian M. The analysis of multivariate longitudinal data: a review. Stat Methods Med Res. 2014;23(1):42-59. doi:10.1177/0962280212445834

Viles WD, Madan JC, Li H, Karagas MR, Hoen AG. Information content of high-order association of the human gut microbiota network, Annals of Applied Statistics. In press.

Zhou J, Deng Y, Luo F, He Z, Tu Q, Zhi X. Functional molecular ecological networks. mBio. 2010 Oct 5;1(4):e00169-10. doi: 10.1128/mBio.00169-10. PMID: 20941329; PMCID: PMC2953006.

Zhou J., Viles W.D., Lu B. et al. Identification of microbial interaction network: zero-inflated latent Ising model based approach. BioData Mining 13, 16 (2020). https://doi.org/10.1186/s13040-020-00226-7

Zhou J, Hoen AG, Mcritchie S, Pathmasiri W, Viles WD, Nguyen QO, Madan JC, Dade E, Karagas MR, Gui J, Information enhanced model selection for Gaussian graphical model with application to metabolomic data, Biostatistics, 2021;, kxab006, https://doi.org/10.1093/biostatistics/kxab006