Microbial Interaction Network Estimation via Bias-Corrected Graphical Lasso

Duo Jiang, Thomas Sharpton, Yuan Jiang*

February 13, 2020

*Duo Jiang is an Assistant Professor at the Department of Statistics, Oregon State University, Corvallis, Oregon 97331, USA (Email: jiangd@stat.oregonstate.edu). Thomas Sharpton is an Assistant Professor at the Department of Microbiology and Department of Statistics, Oregon State University, Corvallis, Oregon 97331, USA (Email: Thomas.Sharpton@oregonstate.edu). Yuan Jiang is the corresponding author. He is an Associate Professor at the Department of Statistics, Oregon State University, Corvallis, Oregon 97331, USA (Email: yuan.jiang@stat.oregonstate.edu). The authors' research is supported in part by National Institutes of Health grant R01 GM126549.

Abstract

With the increasing availability of microbiome 16S data, network estimation has become a useful approach to studying the interactions between microbial taxa. Network estimation on a set of variables is frequently explored using graphical models, in which the relationship between two variables is modeled via their conditional dependency given the other variables. Various methods for sparse inverse covariance estimation have been proposed to estimate graphical models in the high-dimensional setting, including graphical lasso. However, current methods do not address the compositional count nature of microbiome data, where abundances of microbial taxa are not directly measured, but are reflected by the observed counts in an error-prone manner. Adding to the challenge is that the sum of the counts within each sample, termed "sequencing depth", is an experimental technicality that carries no biological information but can vary drastically across samples. To address these issues, we develop a new approach to network estimation, called BC-GLASSO (bias-corrected graphical lasso), which models the microbiome data using a logistic normal multinomial distribution with the sequencing depths explicitly incorporated, corrects the bias of the naive empirical covariance estimator arising from the heterogeneity in sequencing depths, and builds the inverse covariance estimator via graphical lasso. We demonstrate the advantage of BC-GLASSO over current approaches to microbial interaction network estimation under a variety of simulation scenarios. We also illustrate the efficacy of our method in an application to a human microbiome data set.

1 Introduction

Microorganisms are ubiquitous in nature and responsible for managing key ecosystem services (Arrigo, 2004). For example, microbes that colonize the human gut play an important role in homeostasis and disease (Mazmanian et al., 2008; Kamada et al., 2013; Kohl et al., 2014). To better reveal the underlying role microorganisms play in human diseases requires a thorough understanding of how microbes interact with one another. The study of microbiome interactions frequently relies on DNA sequences of taxonomically diagnostic genetic markers (e.g., 16S rRNA), the count of which can then be used to represent the abundance of Operational Taxonomic Units (OTU's, a surrogate for microbial species) in a sample.

The OTU abundance data possess a few important features in nature. First, the data are represented as discrete counts of the 16S rRNA sequences. Second, the data are compositional because the total count of sequences per sample is predetermined by how deeply the sequencing is conducted, a concept named sequencing depth. The OTU counts only carry information about the relative abundances of the taxa instead of their absolute abundances. In addition, the sequencing depth can vary drastically across samples. Last, the OTU data are high-dimensional in nature, as it is likely that the number of OTU's are far more than the number of samples in a biological experiment.

When such data are available, interactions among microbiota can be inferred through correlation analysis (Faust and Raes, 2012). Specifically, if the relative abundances of two microbial taxa are statistically correlated, then it is inferred that they interact on some level. More recent statistical developments have started to take the compositional feature into account and aim to construct sparse networks for the absolute abundances instead of relative abundances. For example, SparCC (Friedman and Alm, 2012), CCLasso (Fang et al., 2015), and REBACCA (Ban et al., 2015) use either an iterative algorithm or a global optimization procedure to estimate the correlation network of all species' absolute abundances while imposing a sparsity constraint on the network.

All the above methods are built upon the marginal correlations between pairs of microbial taxa, and they could lead to spurious correlations that are caused by confounding factors such as other taxa in the same community. Alternatively, interactions among taxa can be modeled through their conditional dependencies given the other taxa, which can eliminate the detection of spurious correlations. In an ideal setting, the Gaussian graphical models are a useful approach to studying the conditional dependency, in which the data are modeled through a multivariate normal distribution and the conditional dependency is determined by the nonzero entries of its inverse covariance matrix. Graphical lasso is a commonly used method to estimate sparse inverse covariance matrix for high-dimensional data under the Gaussian graphical models (Yuan and Lin, 2007; Friedman et al., 2008). However, both the count nature and the compositional features of the microbiome abundance data result in violations of the multivariate normality assumption.

SPIEC-EASI is a popular method for estimating a microbial interaction network that is represented by a sparse inverse covariance matrix between the abundances of species (Kurtz et al., 2015). It is a two-step procedure that first performs a central log-ratio (*clr*) transformation on the observed counts (Aitchison, 1986) and then applies graphical lasso to the transformed abundances. As noted in Kurtz et al. (2015), the *clr*-transformed abundances add up to zero, which leads to a singular covariance matrix and thus an ill-posed problem for estimating its inverse. To overcome this difficulty, SPIEC-EASI treats the covariance matrix of the *clr*-transformed abundances as an approximation to that of the log-transformed abundances that is no longer singular. Therefore, the second graphical lasso step is treated as estimating the well-defined inverse covariance matrix of the log-transformed abundances instead of the above-mentioned ill-posed problem. In other words, SPIEC-EASI is built upon the approximation of two covariance matrices, and thus lacks a clear objective function.

More recently, several other methods have been proposed to infer a microbial network, including gCoda (Fioravanti et al., 2018), CD-trace (Yuan et al., 2019) and SPRING (Yoon et al., 2019). However, existing methods for inverse covariance estimation including these methods and SPIEC-EASI do not properly account for two related features intrinsic to microbiome data: (a) the data are compositional counts in nature, and (b) sequencing depth is finite and varies from sample to sample. In microbiome research, a common strategy to tackle uneven sequencing depths is rarefaction, in which data on samples with higher sequencing depths are thinned by randomly subsampling from the observed counts so that the sequencing depths are the same in the rarefied data. However, this is known to amount to substantial loss of data (McMurdie and Holmes, 2014). Another widely-used practice in microbiome data analysis, also adopted albeit implicitly in SPIEC-EASI, is to simply discard the sequence depth by converting the count data directly to compositional proportions as a proxy for the true relative abundances in a sample. However, it relies on the assumption that the estimated proportion of a taxon in a sample is equal to its true value and ignores the uncertainty of the proportion estimates as reflected by the sampling variance of these estimates. Therefore, this approach does not adequately account for the variation in the microbial counts and has been reported to result in excessive false positives in differential abundance analysis of microbiome data (McMurdie and Holmes, 2014).

In this paper, we show, in the context of covariance estimation, that the proportion-based approach leads to substantial bias in the estimator, which can deteriorate the accuracy of the inferred interaction network. To address this challenge, we quantify the bias by directly modeling the compositional count data. We develop BC-GLASSO (bias-corrected graphical lasso), a method for inverse covariance estimation in microbiome data, which accounts for the compositional count nature of microbiome data and embraces the heterogeneous sequencing depths.

BC-GLASSO is a two-step procedure similar to SPIEC-EASI but possessing key distinctions. First, BC-GLASSO is built upon the logistic normal multinomial distribution that is commonly applied to model compositional count data (Aitchison, 1986; Billheimer et al., 2001; Xia et al., 2013), and thus has a clear objective function. This is a hierarchical model that models the compositional counts using a multinomial distribution and hierarchically the multinomial probabilities using a logistic normal distribution. Compared to SPIEC-EASI, the true covariance matrix is defined on the additive log-ratio (*alr*) transformed multinomial probabilities instead of the *clr*-transformed abundances, with the benefit of being positive-definite and possessing a well-defined inverse matrix. Second, we show that the naive estimator of the true covariance matrix, which is the sample covariance matrix based on the *alr*-transformed abundances, has estimation bias in this hierarchical model. The bias can be approximated by a term that is inversely proportional to the sequencing depths. Last, motivated by the form of the estimation bias, we propose a bias correction procedure by accounting for and, in fact, taking advantage of the heterogeneous sequencing depths. The bias-corrected estimator of the true covariance matrix is easy to compute because it can be written as a weighted average of sample-specific covariance matrix estimators based on the *alr*-transformed abundances. Finally, we apply graphical lasso to estimate a sparse inverse covariance matrix based on this bias-corrected estimator. The non-zero entries in this sparse inverse covariance matrix are interpreted to represent an edge between the associated taxa in a microbial interaction network.

The rest of the paper is organized as follows. In Section 2, we will describe the BC-GLASSO method by introducing the logistic normal multinomial model for the compositional counts, approximating the estimation bias of the naive estimator of the desired covariance matrix, and correcting its estimation bias. In Section 3, we will evaluate the performance of BC-GLASSO via simulation studies and compare it with SPIEC-EASI. We will show that BC-GLASSO performs better in terms of reducing the estimation bias for the covariance matrix and detecting the edges in the microbial interaction networks more accurately. In Section 4, we report a real data application, in which we compare the performance of BC-GLASSO and SPIEC-EASI when applied to the data from the American Gut Project (McDonald et al., 2018). Section 5 concludes this paper with some discussion. Some details for the theoretical derivations in Section 2 are presented in the Appendix.

2 Bias-Corrected Graphical Lasso

2.1 Data and model

Consider an OTU abundance data set with n independent samples, each of which composes observed counts of $K+1$ taxa, denoted by $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,K+1})'$ for the i-th sample, $i = 1, \ldots, n$. Due to the compositional property of the data, the total count of all taxa for each sample i is a fixed number, denoted by M_i. Naturally, a multinomial distribution is imposed on the observed counts:

$$\mathbf{X}_i | \mathbf{P}_i = \mathbf{p}_i \sim \text{Multinomial}(M_i; p_{i,1}, \ldots, p_{i,K+1}), \tag{1}$$

where $\mathbf{p}_i = (p_{i,1}, \ldots, p_{i,K+1})'$ represents the sample-specific multinomial probabilities for individual taxa satisfying that $p_{i,1} + \cdots + p_{i,K+1} = 1$.

To model the variability of the multinomial probabilities in the population, we build a logistic normal distribution on $\mathbf{p}_i$. We first choose one taxon, without loss of generality the $(K+1)$-st taxon, as a reference for all the other taxa and then apply the additive log-ratio (*alr*) transformation (Aitchison, 1986) on the multinomial probabilities:

$$Z_{i,k} = \log\left(\frac{P_{i,k}}{P_{i,K+1}}\right), \; i = 1, \ldots, n, \; k = 1, \ldots, K. \tag{2}$$

Let $\mathbf{Z}_i = (Z_{i,1}, \ldots, Z_{i,K})'$ and further assume that they follow an i.i.d. multivariate normal distribution

$$\mathbf{Z}_i \overset{iid}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\Sigma}),\ i = 1, \ldots, n, \tag{3}$$

where $\boldsymbol{\mu}$ is the mean, and $\boldsymbol{\Sigma}$ is the covariance matrix.

The above model in (1)–(3), known as a logistic normal multinomial model, is a hierarchical model with two levels. The multinomial distribution is imposed on the compositional counts, which is the distribution of the observed data given the multinomial probabilities. In addition, the logistic normal distribution is imposed on the multinomial probabilities as a prior distribution.

The logistic normal multinomial model has been applied to microbiome data to detect covariates that are associated with differential microbial taxa (Xia et al., 2013). The goal of this paper, however, is to infer interactions between microbial taxa. To this end, we set $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ to be the inverse covariance matrix or the precision matrix. $\boldsymbol{\Omega}$ is the parameter of interest whose non-zero entries encode the conditional dependencies between $Z_{i,1}, \ldots, Z_{i,K}$, which are interpreted as edges in the microbial interaction network. Our objective is to find a sparse estimator of the inverse covariance matrix $\boldsymbol{\Omega}$ based on the observed data $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$.

2.2 Naive estimation

A naive approach to estimating $\boldsymbol{\Omega}$ is a two-step procedure. First, one can estimate $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ from the multinomial distribution by applying the same *alr* transformation on the counts as in

$$\hat{Z}_{i,k} = \log\left(\frac{X_{i,k}}{X_{i,K+1}}\right),\ i = 1, \ldots, n,\ k = 1, \ldots, K, \tag{4}$$

and then apply graphical lasso directly on $\hat{\mathbf{Z}}_1, \ldots, \hat{\mathbf{Z}}_n$ by treating them as surrogates for $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$:

$$\ell(\boldsymbol{\Omega}) = -\log[\det(\boldsymbol{\Omega})] + \mathrm{tr}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega}) + \lambda\|\boldsymbol{\Omega}\|_1, \tag{5}$$

where $\hat{\boldsymbol{\Sigma}}$ is the sample covariance matrix of $\hat{\mathbf{Z}}_1, \ldots, \hat{\mathbf{Z}}_n$ and λ is a tuning parameter.

This naive estimation shares the same spirit as SPIEC-EASI (Kurtz et al., 2015) except that the *alr* transformation is used in (4) but SPIEC-EASI uses the central log-ratio (*clr*) transformation

$$\hat{Z}_{i,k} = \log\left(\frac{X_{i,k}}{g(\mathbf{X}_i)}\right),\ i = 1,\ldots,n,\ k = 1,\ldots,K+1,$$

where $g(\mathbf{X}_i)$ is the geometric mean of the counts $X_{i,1},\ldots,X_{i,K+1}$. As noted in Kurtz et al. (2015), the *clr* transformation results in a singular covariance matrix for the transformed data, and thus the non-existence of the inverse covariance matrix. Nonetheless, Kurtz et al. (2015) argued that this covariance matrix is an approximation of the covariance matrix of the logged counts $\log(X_{i,1}),\ldots,\log(X_{i,K+1})$ and they applied graphical lasso to this covariance matrix directly. Therefore, SPIEC-EASI is built upon the approximation of two covariance matrices, and thus lacks a clear objective function. In this paper, we focus on the *alr* transformation in (4) instead of the *clr* transformation and call the resultant estimator of $\mathbf{\Omega}$ from (5) the naive estimator.

In the logistic normal multinomial model in (1)–(3), the inverse covariance matrix is defined on the true parameters $\mathbf{Z}_1,\ldots,\mathbf{Z}_n$ but not their estimators $\hat{\mathbf{Z}}_1,\ldots,\hat{\mathbf{Z}}_n$. The naive estimation treats $\hat{\mathbf{Z}}_1,\ldots,\hat{\mathbf{Z}}_n$ as known, which ignores the variation of $\hat{\mathbf{Z}}_1,\ldots,\hat{\mathbf{Z}}_n$ as the estimators of $\mathbf{Z}_1,\ldots,\mathbf{Z}_n$. In Section 2.3, we will show that the naive estimator has an estimation bias due to the ignorance of the variation of $\hat{\mathbf{Z}}_1,\ldots,\hat{\mathbf{Z}}_n$.

2.3 Estimation bias

In this subsection, we investigate how the naive sample covariance matrix $\hat{\mathbf{\Sigma}}$ based on $\hat{\mathbf{Z}}_1,\ldots,\hat{\mathbf{Z}}_n$ estimates the true covariance matrix $\mathbf{\Sigma}$ in (3). In particular, we evaluate the estimation bias of each element in the covariance matrix separately.

For each pair (k,l) with $1 \le k,l \le K$, let $\sigma_{kl} = \mathrm{Cov}(Z_{i,k}, Z_{i,l})$ be the true covariance between $Z_{i,k}$ and $Z_{i,l}$. Notice that σ_{kl} does not depend on i because $\mathbf{Z}_1,\ldots,\mathbf{Z}_n$ share the same distribution. The naive estimator of σ_{kl} is the sample covariance

$$\hat{\sigma}_{kl} = \frac{1}{n}\sum_{i=1}^{n}\hat{\sigma}_{i,kl} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Z}_{i,k} - \hat{Z}_{\cdot,k})(\hat{Z}_{i,l} - \hat{Z}_{\cdot,l}), \tag{6}$$

where $\hat{\sigma}_{i,kl} = (\hat{Z}_{i,k} - \hat{Z}_{\cdot,k})(\hat{Z}_{i,l} - \hat{Z}_{\cdot,l})$ and $\hat{Z}_{\cdot,k} = \frac{1}{n}\sum_{i=1}^{n}\hat{Z}_{i,k}$ for $k = 1,\ldots,K$. In (6), it is seen that the sample covariance $\hat{\sigma}_{kl}$ is the arithmetic mean of $\hat{\sigma}_{1,kl},\ldots,\hat{\sigma}_{n,kl}$, the corresponding contributions from each sample. In the following, we will argue that $\hat{\sigma}_{i,kl}$ is biased as an estimator of σ_{kl} and so is $\hat{\sigma}_{kl}$.

When M_i is large, the Taylor's expansion of $X_{i,k}/M_i$ at its conditional mean $P_{i,k}$ gives the following approximation by ignoring higher-order terms

$$\hat{Z}_{i,k} \approx Z_{i,k} + \left(\frac{X_{i,k}}{M_i P_{i,k}} - \frac{X_{i,K+1}}{M_i P_{i,K+1}} \right) - \frac{1}{2} \left[\left(\frac{X_{i,k}}{M_i P_{i,k}} - 1 \right)^2 - \left(\frac{X_{i,K+1}}{M_i P_{i,K+1}} - 1 \right)^2 \right]. \qquad (7)$$

A direct evaluation leads to the following approximations

$$\mathrm{E}(\hat{Z}_{i,k}) \approx \mu_k - \frac{C_k}{2M_i} \quad \text{and} \quad \mathrm{E}(\hat{Z}_{i,k}\hat{Z}_{i,l}) \approx \sigma_{kl} + \mu_k\mu_l + \frac{C_{kl}}{M_i}, \qquad (8)$$

where $C_k = \mathrm{E}(P_{1,k}^{-1} - P_{1,K+1}^{-1})$ and $C_{kl} = \mathrm{E}(P_{1,K+1}^{-1}) - \frac{1}{2}\mathrm{Cov}(Z_{1,k}, P_{1,l}^{-1} - P_{1,K+1}^{-1}) - \frac{1}{2}\mathrm{Cov}(Z_{1,l}, P_{1,k}^{-1} - P_{1,K+1}^{-1})$ are quantities that do not depend on i. Plugging (8) to $E(\hat{\sigma}_{i,kl})$ leads to the following approximation of its estimation bias when M_i and n are both large

$$E(\hat{\sigma}_{i,kl}) \approx \sigma_{kl} + \frac{C_{kl}}{M_i}. \qquad (9)$$

For details of the above derivations, we refer to the Appendix. The result in (9) implies that $\hat{\sigma}_{i,kl}$ has an approximate bias with the order of M_i^{-1} as an estimator of σ_{kl}, ignoring higher-order terms. In addition, as the arithmetic mean of $\hat{\sigma}_{i,kl}$, the naive sample covariance $\hat{\sigma}_{kl}$ is also approximately biased with the order of $\frac{1}{n}\sum_{i=1}^{n} M_i^{-1}$ for the bias term, when all M_i's and n are large.

The expression in (9) has a similar form to a simple linear regression if we treat $\hat{\sigma}_{1,kl}, \ldots, \hat{\sigma}_{n,kl}$ as the responses and $M_1^{-1}, \ldots, M_n^{-1}$ as the explanatory variables. In this linear regression, σ_{kl} serves as the intercept and C_{kl} serves as the slope. This observation motivates us to develop a bias correction procedure based on fitting such a simple linear regression in Section 2.4.

2.4 Bias correction and graphical lasso

We fit a simple linear regression of the responses $\hat{\sigma}_{1,kl}, \ldots, \hat{\sigma}_{n,kl}$ to the explanatory variables $M_1^{-1}, \ldots, M_n^{-1}$:

$$\hat{\sigma}_{i,kl} \sim \beta_0 + \beta_1 \frac{1}{M_i},\ i = 1, \ldots, n$$

and use the least-squares estimator of the intercept β_0, denoted by $\tilde{\sigma}_{kl}$, to estimate σ_{kl}. It is not hard to show that

$$\tilde{\sigma}_{kl} = \frac{1}{n} \sum_{i=1}^{n} (1 + \delta_i)\, \hat{\sigma}_{i,kl}, \tag{10}$$

where

$$\delta_i = \frac{\|\mathbf{M}^{-1}\|_1}{\|\mathbf{M}^{-1}\|_2^2 - \|\mathbf{M}^{-1}\|_1^2/n} \left(\frac{\|\mathbf{M}^{-1}\|_1}{n} - \frac{1}{M_i} \right),$$

where $\mathbf{M}^{-1} = (M_1^{-1}, \ldots, M_n^{-1})'$, and $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the L_1 and L_2 norm of a vector, respectively.

Compared to the naive estimator $\hat{\sigma}_{kl}$ that is an arithmetic mean of $\hat{\sigma}_{1,kl}, \ldots, \hat{\sigma}_{n,kl}$, the bias corrected estimator $\tilde{\sigma}_{kl}$ is a weighted mean. It is seen that when the sample i has a higher sequencing depth M_i, δ_i will be larger and so is the weight, which agrees with the intuition that a higher sequencing depth gives more accuracy in estimating its compositional probabilities. In addition, the fact that $\sum_{i=1}^n \delta_i = 0$ implies that the sum of all the weights still add up to one as same as in the naive estimator.

Fig. 1 presents a scatter plot of the estimated and true covariances in the (1,2)-entry of the covariance matrix from a simulation study. If an estimator has no bias, the points on the scatter plot should lie around the straight line which indicates the equality of the quantities on both axes. It is obvious that the naive estimator has a substantial bias (left panel) and the bias correction procedure is effective in removing the bias (right panel). This can also be justified by evaluating $\mathrm{E}(\tilde{\sigma}_{kl})$ by combining (9) and (10), which turns out to be approximately unbiased as

$$E(\tilde{\sigma}_{i,kl}) \approx \sigma_{kl}.$$

In another word, the bias-corrected estimator $\tilde{\sigma}_{kl}$ is approximately unbiased when all sequencing depths and the sample size are large. In Fig. 1, we also notice that the variance of the bias-corrected estimator $\tilde{\sigma}_{kl}$ is slightly higher than that of the naive estimator $\hat{\sigma}_{kl}$. In this particular simulation, the sample variance of $\tilde{\sigma}_{kl}$ is approximately twice of the sample variance of $\hat{\sigma}_{kl}$.

Insert Fig. 1 here.

With the bias-corrected estimator of the covariance matrix $\tilde{\mathbf{\Sigma}}$ whose (k, l)-entry being $\tilde{\sigma}_{kl}$ for $1 \leq k, l \leq K$, we can apply graphical lasso to achieve a sparse estimator of the inverse covariance matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ as in

$$\ell(\mathbf{\Omega}) = -\log[\det(\mathbf{\Omega})] + \mathrm{tr}(\tilde{\mathbf{\Sigma}}\mathbf{\Omega}) + \lambda\|\mathbf{\Omega}\|_1. \tag{11}$$

Note that the only difference of (11) from (5) is the replacement of the naive estimator $\hat{\mathbf{\Sigma}}$ by the bias-corrected estimator $\tilde{\mathbf{\Sigma}}$. Therefore, we call this approach bias-corrected graphical lasso (BC-GLASSO) and the resultant inverse covariance matrix estimator the BC-GLASSO estimator.

In the following sections, we will apply BC-GLASSO on simulated data and real data to evaluate its performance by assessing its estimation unbiasedness for the covariance matrix and its identification accuracy for the interaction network.

3 Simulation Studies

We perform simulation studies under a variety of settings to assess the effectiveness of bias correction and the accuracy of network identification using BC-GLASSO and to compare its performance with SPIEC-EASI. Note that in addition to adopting the naive approach described in Section 2.2, SPIEC-EASI also uses the *clr* transformation rather than the *alr* transformation in BC-GLASSO. For a fair comparison and to highlight the impact of the proposed bias correction technique enabled by more careful modeling of the data, we will adopt the *alr* transformation in our implementation of SPIEC-EASI instead of its default *clr* transformation. We also implement CD-trace and gCoda for comparison.

3.1 Simulation settings

We consider four types of network structures: the random-edge, cluster, hub, and scale-free networks (Fig. 2). First, in the random-edge network, each pair of nodes, independently of other pairs, has a probability of 0.3 to be connected by an edge. In its corresponding inverse covariance matrix $\mathbf{\Omega}$, ω_{kl} is 1 if nodes k and l are connected and 0 otherwise, while ω_{kk} is a constant in k that controls the condition number of $\mathbf{\Omega}$ at 100. Second, in a cluster network, the nodes are evenly partitioned into 2 disjoint groups of the same size. Each group form a cluster that is interconnected as a random-edge network with the connection probability 0.3. Third, similar to the cluster network, the nodes in a hub network are also evenly partitioned into 2 disjoint groups of the same size and each group has a center to which all the other nodes within the same group are connected to. Finally, in a scale-free network, the distribution of degrees (the number of connections each node has to other nodes) follows a power law. The cluster, hub, and scale-free networks as well as their respective inverse covariance matrices are generated using the `Huge` package in R (Albert and Barabási, 2002; Zhao et al., 2012). In this package, we set the off-diagonal element of $\mathbf{\Omega}$ to be $v = 0.3$ for the cluster network and $v = 0.03$ for the hub and scale-free networks. Throughout our simulations, we fix the number of nodes to be $K = 50$ for all four types of networks.

Insert Fig. 2 here.

For each network structure, we simulate microbiome compositional counts on $n = 500$ samples with heterogeneous sequencing depths given by $\mathbf{M} = (M_1, \cdots, M_n)'$. We consider three settings for $\mathbf{M}$: (M1) half of the M_i's are $K/2.5$ and the other half are $40K$, (M2) each M_i is independently drawn from the uniform distribution from $K/2.5$ to $40K$, and (M3) the M_i's are generated from the real sequencing depths in the 16S data from the American Gut Project (see Section 4). Specifically, for setting (M3), after removing rare OTU's (average relative abundance $< 0.01\%$) in the real data, we compute the total reads of each sample based on a randomly selected set of $K + 1$ taxa. Then, after removing samples whose total reads are below $K+1$, we randomly draw 500 samples from the rest and use their total reads as $\mathbf{M}$. In summary, settings (M1) and (M2) contrast cases with high and low heterogeneity in sequencing depth, while setting (M3) tries to mimic the real situation.

Given the inverse covariance matrix $\mathbf{\Omega}$ for each network structure, we independently draw $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ from the multivariate normal distribution given by $N(\boldsymbol{\mu}, \mathbf{\Omega}^{-1})$, where $\boldsymbol{\mu} = (\mu_0, \cdots, \mu_0)'$. In general, with other factors held fixed, the greater μ_0 is, the rarer the reference taxon tends to be. In our simulations, we use $\mu_0 = \log(4/K) < 0$, which implies that the reference taxon is on average more abundant than the other taxa. Given $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ and the sequencing depths $\mathbf{M}$ generated as described in the previous paragraph, the compositional counts are generated based on the multinomial distribution in (1) and (2). The simulated count data may include zero counts. In order to perform the log-ratio transformation to obtain the $\hat{\mathbf{Z}}_i$'s, we add to each count $X_{i,k}$ a small positive number equal to $\hat{p}_k(K+1)$, where $\hat{p}_k = n^{-1}\sum_{i=1}^{n} X_{i,k}/M_i$ is the estimated mean relative abundance of taxon k. This allows a zero count to be replaced by a positive number whose value depends on the relative abundance of the associated taxon in the other samples for which its observed abundance is non-zero. For each simulation setting, the process described above is independently repeated to create 100 replicates.

3.2 Simulation results

First, to assess the effectiveness of BC-GLASSO in correcting the estimation bias in the covariance matrix, we compare the estimated covariances to their corresponding true values across simulation replicates to obtain the empirical bias and mean squared error (MSE) separately for each off-diagonal entry in the covariance matrix. Table 1 summarizes the results by averaging the absolute values of the empirical bias and the MSE values across all pairs of taxa. In addition, the entry-wise bias values are visualized in heatmaps (see Fig. S1-S4 in the Supplemental Materials).

To assess the accuracy with which a method is able to recover the interaction network, we compare the true network with the inferred network obtained by joining pairs of nodes with non-zero entries in the estimated inverse covariance matrix. The true positive rate is defined to be the frequency with which an edge in the true network is present in the inferred network, and the false positive rate is defined to be the frequency with which an edge not present in the true network is identified in the inferred network. We implement BC-GLASSO,

Table 1: Simulation results. Bias: average of the absolute values of the empirical bias across all pairs of taxa; MSE: average of the MSE values across all pairs of taxa.

		(M1)		(M2)		(M3)	
		Bias	MSE	Bias	MSE	Bias	MSE
Random Edge	SPIEC-EASI	6.79E-4	2.34E-3	1.27E-2	6.30E-4	1.36E-2	6.79E-4
	BC-GLASSO	5.71E-4	4.75E-4	2.94E-3	5.25E-4	4.21E-3	5.71E-4
Cluster	SPIEC-EASI	6.69E-2	8.27E-3	2.26E-2	2.68E-3	2.07E-2	2.56E-3
	BC-GLASSO	3.84E-2	5.95E-3	7.51E-3	2.88E-3	5.93E-3	3.10E-3
Hub	SPIEC-EASI	4.93E-2	3.77E-3	1.41E-2	2.00E-3	1.28E-2	1.99E-3
	BC-GLASSO	3.61E-2	5.74E-3	4.19E-3	2.86E-3	4.31E-3	2.83E-3
Scale Free	SPIEC-EASI	5.25E-2	4.12E-3	1.47E-2	1.99E-3	1.18E-2	1.95E-3
	BC-GLASSO	3.76E-2	5.84E-3	4.54E-3	2.83E-3	4.28E-3	2.99E-3

SPIEC-EASI, CD-trace and gCoda with a range of values for the tuning parameter, which allows us to plot the true positive rate of each method against its false positive rate in an ROC curve. For CD-trace and gCoda, because the true network has one less node than the dimension of the estimated inverse covariance matrix, we build the estimated network based on the K-dimensional submatrix on the top left by excluding the last dimension. The ROC curves from different simulation settings are included in Fig. 3.

Insert Fig. 3 Here.

From Table 1, we find that BC-GLASSO effectively reduces the overall bias by up to 2 orders of magnitude. We note that under some scenarios, the bias reduction comes at the cost of a moderate increase in MSE. This is due to a potential increase in the variance of the bias-corrected estimator as compared to the naive estimator. In addition, as demonstrated in Fig. 3, BC-GLASSO outperforms the naive procedure in recovering the interaction network across all network structures and sequencing depth settings in our simulations. More specifically, at a fixed true positive rate, BC-GLASSO consistently maintains a lower false positive rate than SPIEC-EASI, and at a fixed false positive rate, BC-GLASSO consistently attains a higher true positive rate. BC-GLASSO exhibits a greater advantage over SPIEC-EASI in a setting with high heterogeneity in the sequencing depths (M1) than in a setting with low heterogeneity (M2). Substantial improvement in network recovery is achieved by BC-

GLASSO when the sequencing depths are obtained mimicking the real situation in setting (M3). For example, for the scale-free network, BC-GLASSO, when compared to SPIEC-EASI, reduces the false positive rate from 24.7% to 15.4% with a fixed true positive rate of 90%. For a random-edge network, BC-GLASSO increases the true positive rate from 44.8% to 59.0% with a fixed false positive rate of 20%. The comparison with CD-trace shows that BC-GLASSO either outperforms or achieves very similar performance to CD-trace in most of the settings, with the exception of a hub network with high sequencing depth heterogeneity (M1), for which CD-trace is slightly better than BC-GLASSO. However, BC-GLASSO yields substantial improvement over CD-trace in several settings including the cluster network with M1. Overall, we conclude that BC-GLASSO compares favorably with CD-trace. Moreover, the comparison with gCoda indicates that the performance of gCoda is dominated by that of CD-trace and BC-GLASSO.

In summary, BC-GLASSO is effective in reducing bias in the estimation of the covariance matrix compared to SPIEC-EASI. In some scenarios, BC-GLASSO can yield a higher MSE due to the inflation of the estimation variance. However, in all of our simulation scenarios, BC-GLASSO always outperforms SPIEC-EASI in terms of the accuracy in the recovering the interaction network represented by the estimated inverse covariance matrix. The overall performance of BC-GLASSO also surpasses that of CD-trace.

4 Real Data Analysis

We illustrate the use of BC-GLASSO with an analysis of 16S data from the American Gut Project (AGP) (McDonald et al., 2018) to infer the interaction network between microbial taxa in the human gut. We focus on the data collected on 3,679 stool samples and remove from the data set samples collected from other body sites.

To better capture the variation in the composition of human microbiota, it is helpful to take into account subgroup structure of the population that may exhibit fundamentally different biological properties. Recent research has found evidence that while the gut microbiome takes on smooth gradients of compositional diversity across individuals (Koren et al. 2013), human populations can generally be stratified into two main clusters, referred to as “enterotypes”, based on the abundance of specific taxa (Arumugam et al., 2011). These enterotypes are compositionally and potentially functionally distinct (Costea et al., 2018). In our analysis, we propose to estimate the microbial interaction network separately for each enterotype, which helps minimize the detection of spurious interactions due to confounders related to population stratification and enhances our ability to identify biologically relevant interactions. Recent studies have shown the existence of two common enterotypes, including one marked by a high relative abundance of *Bacteroides* and the other by a high relative abundance of *Prevotella* (Gorvitovskaia et al., 2016), and that *Prevotella*-to-*Bacteroides* ratio (P/B ratio) can be used to effectively classify humans to these subpopulations (Roager et al., 2014). In our analysis of the AGP data, we analyze two groups of samples separately: the P group which includes samples whose P/B ratio is in the upper 25%, and the B group includes samples whose P/B ratio is in the lower 25%.

Insert Fig. 4 Here.

We conduct our analysis on the genus level and aggregate the counts for OTU’s that belong to the same genus. OTU’s the do not have the genus-level taxonomic information are aggregated into a pseudo-genus, which we will refer to as the “unlabeled genus.” We filter the data to remove samples with very low sequencing depths and genera that are highly sparse. More specifically, samples with total reads smaller than 100 are removed from the analysis. Separately for the two groups based on P/B ratio, we remove genera with zero abundance for over 95% of the subjects within a group. The resulting data set has 786 subjects and 143 genera in the P group, and 864 subjects and 142 genera in the B group.

To select the reference taxon, we aim to find a genus whose abundance remains stable across samples. To this end, we evaluate the inter-sample dispersion of the relative abundance of each taxon and find the taxon whose relative abundance is the least dispersed. More specifically, we first calculate each genus’ relative abundance in a sample by taking the ratio of its observed count to the total read count for the sample. Then, we measure the dispersion of a taxon’s relative abundance by the coefficient of variation, defined as the standard deviation of the relative abundance across samples divided by its mean. The genus that minimizes the coefficient of variation is taken as the reference taxon. This criterion is applied separately for the P group and the B group, for both of which the unlabeled genus is chosen to be the reference. This reference has a non-zero count for over 95% of the samples in both groups.

We apply BC-GLASSO and SPIEC-EASI to estimate the inverse covariance matrix separately for the two groups. For both methods, the value of the tuning parameter λ is selected using a grid search based on BIC. In Fig. 5–7, we compare the results based on BC-GLASSO and SPIEC-EASI for the P group. The comparison between the two methods is qualitatively similar for the B group and we show the results in Fig. S5–S7 in the Supplemental Materials. Fig. 5 visualizes the estimated correlation matrices based on the two methods. SPEIC-EASI has resulted in an estimated correlation matrix wherein positive correlations overwhelmingly dominate negative ones. In contrast, the correlation matrix based on BC-GLASSO has led to a greater number of negative correlations.

Insert Fig. 5 Here.

To visualize the microbial interactions identified by a method, we show the heatmap visualizing the estimated inverse covariance matrix, where a non-zero entry corresponds to an edge in the inferred microbial network (Fig. 6). Because an entry in an inverse covariance matrix is connected to the negative of the partial correlation between two variables (Lauritzen, 1996), we show the inverse covariance matrices on the negative scale. Both methods are able to identify two clusters of taxa that are densely connected to each other. The first cluster includes about 14 genera from the family *Enterobacteriaceae*, and the second cluster includes all of the seven genera from the family *[Tissierellaceae]*. Comparing the inverse covariance matrices based on the two methods, however, SPIEC-EASI seems to lead to a background rate of additional non-zero interactions that arise evenly from all families, while for BC-GLASSO the identified interactions align more closely with the taxonomic relationships of the genera, showing a clearer trend of genera from the same family exhibiting similar patterns of interactions.

Insert Fig. 6 Here.

We are further interested in the interactions that are exclusively detected by only one method. In summary, between the 142 genera analyzed for the P group, there are a total of 10,011 potential pairwise interactions. Among these, 938 interactions are identified by both methods, BC-GLASSO detects an additional set of 151 interactions, and SPEIC-EASI detects an additional set of as many as 1,460 interactions (Fig. 7, Panel B). Interestingly, 114 out of the 151 interactions exclusively identified by BC-GLASSO (shown in red in Fig. 7, Panel A) are associated with a small number of genera, all of which are from the family *Enterobacteriaceae*. These genera are found to have extensive interactions with the rest of the microbiome which are captured by BC-GLASSO. The genera include *Pantoea* (44), *Enterobacter* (34), *Plesiomonas* (14), *Klebsiella* (12), and *Erwinia* (10), where the numbers in parentheses indicate how many edges associated with a genus are unique identified by BC-GLASSO. The interactions associated with these genera exclusively identified by BC-GLASSO are listed in Supplemental Table S1. In contrast, the 1,460 interactions that are present only in the network produced by SPIEC-EASI (shown in white in Fig. 7, Panel A) are widespread and distributed across all families, rather than concentrated within specific taxa.

The unique interactions revealed by BC-GLASSO represent important discoveries that can advance follow-up research and potentially impact the development of clinical resources. For example, members of the genera *Enterobacteria*, *Klebsiella*, and *Plesiomonas* include known opportunistic pathogens and pathobionts that can impact the host of the health when highly abundant in the gut (Davin-Regli et al., 2015; Schneditz et al., 2014; Janda et al., 2016). Our identification of additional linkages between members of these genera and other gut taxa can help guide experiments that uncover how other gut taxa impact the success of these organisms in the gut, possibly through competitive exclusion or biocontrol.

Insert Fig. 7 Here.

5 Discussion

It is becoming increasingly recognized that microbiome data have unique characteristics that are known to require tailored statistical methods. With these characteristics in mind, in this paper, we focus on the problem of inferring the interaction network between microbial taxa through the estimation of a sparse inverse covariance estimation in microbiome data. We have highlighted a key disadvantage of the popular proportion-based approach due to the bias that originates from the failure to properly model the abundance counts and to adequately capture the variation in the data . To address this issue, we have developed BC-GLASSO, a model-based method for inverse covariance estimation which directly tackles the compositional count data and exploits the heterogeneity in sequencing depth. Features of BC-GLASSO include: (a) the method is based on a hierarchical model where the technical variation of the count data are modeled using a multinomial distribution and the biological (i.e. inter-sample) variation of microbiome composition is modeled using a logistic normal distribution on the multinomial probabilities; (b) the unevenness in the sequencing depth, which frequently poses a challenge in microbiome data analysis, is not only properly accounted for in our model but also taken advantage of to correct the bias in the estimator; (c) Despite the hierarchical model used in BC-GLASSO, the method remains computationally rapid even on big data sets owing to the linear models underlying the bias correction procedure.

We have demonstrated the advantage of BC-GLASSO relative to a leading approach through simulation studies. In particular, BC-GLASSO consistently outperforms the competing method under a variety of network structures and different setups for the sequencing depths. The strength of BC-GLASSO is manifested by a greater accuracy in the inferred network as well the reduced bias of the covariance estimator. We have also applied BC-GLASSO to infer the microbial interaction network in a data set from the American Gut Project, where BC-GLASSO has detected a group of genera from the *Enterobacteriaceae* family which have extensive interactions with the rest of the microbiome.

In our presentation of BC-GLASSO, the $\mathbf{Z}_i$'s are assumed to follow $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We note that the normality assumption on $\mathbf{Z}_i$ is not essential. In fact, even without specific distributional assumptions on the $\mathbf{Z}_i$'s, all theoretical derivations and properties reported in this paper on the covariance estimator, $\hat{\sigma}_{kl}$, in BC-GLASSO remain valid so long as $\mathbf{Z}_i$'s are assumed to i.i.d. with some mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. The use of graphical lasso in the second step of BC-GLASSO, however, does assume normality of $\mathbf{Z}_i$'s.

One caveat our approach from the perspective of biological interpretation is that the interaction network excludes up to k possible edges that theoretically could exist in the community being studied. Specifically, potential interactions between the reference taxon and all other taxa are not modeled by our method and thus cannot be analyzed and interpreted by users. If users are interested in understanding how particular taxa relate, then they would want to avoid using such taxa as a reference given that our approach specifically excludes such reference taxa from the final interaction network. Without extensive prior biological information, we recommend taking as the reference a taxon whose relative abundance is stable across samples.

Microbiome studies often record metadata on the samples including covariates such as the age, sex, and dietary information of a subject. Some of these covariates have been associated with the abundance of a taxon in the microbiome. It can be helpful for such associations to be accounted for when estimating the microbial interaction network. To this end, a potential approach is to extend GC-GLASSO to incorporate covariates in the hierarchical model. This may be done, for example, by allowing $\boldsymbol{\mu}$ to depend on covariates. This is beyond the scope of this paper but may present worthwhile opportunities for future research.

Although BC-GLASSO is motivated by problems that arise in microbiome research, it can be applied to compositional count data from other types of applications so long as the total count varies substantially across samples. Examples include ecological data on species abundance (Jackson, 1997), where it may be of interest to estimate the ecological relationship between species, and RNA-Seq data, where it may be of interest to infer the regulatory relationship between genes.

6 Figure Legends

Fig. 1 Scatter plot of estimated and true covariances for the (1,2)-entry of the covariance matrix from a simulation study. Results are based on $K = 50$, $n = 500$, setting (M1) for

the sequencing depth as described in the simulation setting, and a true correlation between the first and second taxa ranging from -0.9 to 0.9. Left panel: naive estimator; right panel: bias-corrected estimator.

Fig. 2 The four network structures used in simulation studies. Panel A, random-edge network. Panel B, cluster network. Panel C, hub network; Panel D, scale-free network.

Fig. 3 BC-GLASSO enjoys more accurate network recovery than SPIEC-EASI in simulations. ROC curves for BC-GLASSO and SPIEC-EASI are compared for each simulation setting. Blue: BC-GLASSO. Black: SPIEC-EASI. Each row is a network structure. Each column is a configuration for the sequencing depth.

Fig. 4 Empirical distribution of P/B ratio from the AGP data and the regions in this distribution corresponding to the P group and the B group.

Fig. 5 Estimated correlation matrices for the P group in the AGP data based on BC-GLASSO and SPIEC-EASI. The correlation matrices are obtained based on the estimated covariance matrix used in the two methods. Left: BC-GLASSO. Right: SPIEC-EASI.

Fig. 6 Estimated inverse covariance matrices for the P group in the AGP data based on BC-GLASSO and SPIEC-EASI, on the negative scale. Diagonal entries are omitted. Left: BC-GLASSO. Right: SPIEC-EASI.

Fig. 7 Comparing the microbial networks based on BC-GLASSO and SPIEC-EASI for the P group in the AGP data. Panel A, network difference between two methods. Red - edges identified by BC-GLASSO but not by naive; White - edges identified by naive but not by BC-GLASSO; Orange - edges identified by either both or neither. Panel B, numbers of common and unique edges detected by BC-GLASSO and/or SPIEC-EASI.

Appendix: Justification of Estimation Bias in (9)

For $i = 1, \ldots, n$, taking the Taylor's expansion of $X_{i,k}/M_i$ at its conditional mean $P_{i,k}$ leads to the following high-order terms

$$\log\left(\frac{X_{i,k}}{M_i}\right) \approx \log(P_{i,k}) + \left(\frac{X_{i,k}}{M_i P_{i,k}} - 1\right) - \frac{1}{2}\left(\frac{X_{i,k}}{M_i P_{i,k}} - 1\right)^2.$$

Therefore, $\hat{Z}_{i,k} = \log(X_{i,k}/X_{i,K+1})$ can be written as

$$\hat{Z}_{i,k} \approx Z_{i,k} + \left(\frac{X_{i,k}}{M_iP_{i,k}} - \frac{X_{i,K+1}}{M_iP_{i,K+1}}\right) - \frac{1}{2}\left[\left(\frac{X_{i,k}}{M_iP_{i,k}} - 1\right)^2 - \left(\frac{X_{i,K+1}}{M_iP_{i,K+1}} - 1\right)^2\right],$$

which is also given in (7).

We can approximate $\mathrm{E}(\hat{Z}_{i,k})$ and $\mathrm{E}(\hat{Z}_{i,k}\hat{Z}_{i,l})$ as follows.

$$\begin{aligned}\mathrm{E}(\hat{Z}_{i,k}) &\approx \mathrm{E}(Z_{i,k}) - \frac{1}{2}\mathrm{E}\left[\frac{M_iP_{i,k}(1-P_{i,k})}{M_i^2P_{i,k}^2} - \frac{M_iP_{i,K+1}(1-P_{i,K+1})}{M_i^2P_{i,K+1}^2}\right] \\ &= \mu_k - \frac{1}{2M_i}\mathrm{E}\left(\frac{1}{P_{i,k}} - \frac{1}{P_{i,K+1}}\right),\end{aligned}$$

and $\mathrm{E}(\hat{Z}_{i,k}\hat{Z}_{i,l}) \approx I_1 + I_2 + I_3 + I_4$, where

$$\begin{aligned}I_1 &= E(Z_{i,k}Z_{i,l}) = \mathrm{Cov}(Z_{i,k}, Z_{i,l}) + \mathrm{E}(Z_{i,k})\mathrm{E}(Z_{i,l}) = \sigma_{kl} + \mu_k\mu_l, \\ I_2 &= \mathrm{E}\left[\left(\frac{X_{i,k}}{M_iP_{i,k}} - \frac{X_{i,K+1}}{M_iP_{i,K+1}}\right)\left(\frac{X_{i,l}}{M_iP_{i,l}} - \frac{X_{i,K+1}}{M_iP_{i,K+1}}\right)\right] = \frac{1}{M_i}\mathrm{E}\left(\frac{1}{P_{i,K+1}}\right), \\ I_3 &= -\frac{1}{2}\mathrm{E}\left[Z_{i,k}\left(\frac{X_{i,l}^2}{M_i^2P_{i,l}^2} - \frac{X_{i,K+1}^2}{M_i^2P_{i,K+1}^2}\right)\right] = -\frac{1}{2M_i}\mathrm{E}\left[Z_{i,k}\left(\frac{1}{P_{i,l}} - \frac{1}{P_{i,K+1}}\right)\right], \\ I_4 &= -\frac{1}{2}\mathrm{E}\left[Z_{i,l}\left(\frac{X_{i,k}^2}{M_i^2P_{i,k}^2} - \frac{X_{i,K+1}^2}{M_i^2P_{i,K+1}^2}\right)\right] = -\frac{1}{2M_i}\mathrm{E}\left[Z_{i,l}\left(\frac{1}{P_{i,k}} - \frac{1}{P_{i,K+1}}\right)\right].\end{aligned}$$

In addition, let $\nu_k = \mathrm{E}(\hat{Z}_{\cdot,k})$ for $k = 1, \ldots, K$, then it can be evaluated as

$$\nu_k = \mathrm{E}(\hat{Z}_{\cdot,k}) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(\hat{Z}_{i,k}) = \mu_k - \frac{1}{2n}\sum_{i=1}^{n}\frac{1}{M_i}\mathrm{E}\left(\frac{1}{P_{i,k}} - \frac{1}{P_{i,K+1}}\right).$$

Therefore, $E[(\hat{Z}_{i,k}-\nu_k)(\hat{Z}_{i,l}-\nu_l)]$ can be approximated by

$$\begin{aligned}
&\mathrm{E}[(\hat{Z}_{i,k}-\nu_k)(\hat{Z}_{i,l}-\nu_l)] \\
\approx\ &\mathrm{E}(\hat{Z}_{i,k}\hat{Z}_{i,l})-\nu_k\mathrm{E}(\hat{Z}_{i,l})-\nu_l\mathrm{E}(\hat{Z}_{i,k})+\nu_k\nu_l \\
=\ &\sigma_{kl}+\frac{1}{M_i}\mathrm{E}\left(\frac{1}{P_{i,K+1}}\right)-\frac{1}{2M_i}\mathrm{E}\left[Z_{i,k}\left(\frac{1}{P_{i,l}}-\frac{1}{P_{i,K+1}}\right)\right]-\frac{1}{2M_i}\mathrm{E}\left[Z_{i,l}\left(\frac{1}{P_{i,k}}-\frac{1}{P_{i,K+1}}\right)\right] \\
&+\frac{1}{2M_i}\mu_k\mathrm{E}\left(\frac{1}{P_{i,l}}-\frac{1}{P_{i,K+1}}\right)+\frac{1}{2M_i}\mu_l\mathrm{E}\left(\frac{1}{P_{i,k}}-\frac{1}{P_{i,K+1}}\right) \\
&+\left[\frac{1}{2n}\sum_{i=1}^{n}\frac{1}{M_i}\mathrm{E}\left(\frac{1}{P_{i,k}}-\frac{1}{P_{i,K+1}}\right)\right]\left[\frac{1}{2n}\sum_{i=1}^{n}\frac{1}{M_i}\mathrm{E}\left(\frac{1}{P_{i,l}}-\frac{1}{P_{i,K+1}}\right)\right] \\
\approx\ &\sigma_{kl}+\frac{1}{M_i}\mathrm{E}\left(\frac{1}{P_{i,K+1}}\right)-\frac{1}{2M_i}\mathrm{Cov}\left(Z_{i,k},\frac{1}{P_{i,l}}-\frac{1}{P_{i,K+1}}\right)-\frac{1}{2M_i}\mathrm{Cov}\left(Z_{i,l},\frac{1}{P_{i,k}}-\frac{1}{P_{i,K+1}}\right).
\end{aligned}$$

The above approximation ignores the following term

$$\left[\frac{1}{2n}\sum_{i=1}^{n}\frac{1}{M_i}\mathrm{E}\left(\frac{1}{P_{i,k}}-\frac{1}{P_{i,K+1}}\right)\right]\left[\frac{1}{2n}\sum_{i=1}^{n}\frac{1}{M_i}\mathrm{E}\left(\frac{1}{P_{i,l}}-\frac{1}{P_{i,K+1}}\right)\right],$$

which is small compared to the other terms when all M_i's are large.

Finally, $E(\hat{\sigma}_{i,kl})$ can be written as

$$\begin{aligned}
\mathrm{E}(\hat{\sigma}_{i,kl}) &= \mathrm{E}[(\hat{Z}_{i,k}-\hat{Z}_{\cdot,k})(\hat{Z}_{i,l}-\hat{Z}_{\cdot,l})] \\
&= \mathrm{E}[(\hat{Z}_{i,k}-\nu_k)(\hat{Z}_{i,l}-\nu_l)]-\mathrm{E}[(\hat{Z}_{i,k}-\nu_k)(\hat{Z}_{\cdot,l}-\nu_l)] \\
&\quad -\mathrm{E}[(\hat{Z}_{\cdot,k}-\nu_k)(\hat{Z}_{i,l}-\nu_l)]+\mathrm{E}[(\hat{Z}_{\cdot,k}-\nu_k)(\hat{Z}_{\cdot,l}-\nu_l)].
\end{aligned}$$

By Cauchy-Schwartz's inequality,

$$\begin{aligned}
\mathrm{E}[|(\hat{Z}_{i,k}-\nu_k)(\hat{Z}_{\cdot,l}-\nu_l)|] &\le \sqrt{\mathrm{E}[(\hat{Z}_{i,k}-\nu_k)^2]}\sqrt{\mathrm{Var}(\hat{Z}_{\cdot,l})}; \\
\mathrm{E}[|(\hat{Z}_{\cdot,k}-\nu_k)(\hat{Z}_{i,l}-\nu_l)|] &\le \sqrt{\mathrm{E}[(\hat{Z}_{i,l}-\nu_l)^2]}\sqrt{\mathrm{Var}(\hat{Z}_{\cdot,k})}; \\
\mathrm{E}[|(\hat{Z}_{\cdot,k}-\nu_k)(\hat{Z}_{\cdot,l}-\nu_l)|] &\le \sqrt{\mathrm{Var}(\hat{Z}_{\cdot,k})}\sqrt{\mathrm{Var}(\hat{Z}_{\cdot,l})}.
\end{aligned}$$

Thus, when n is large, these terms can be omitted compared to the first term so that $E(\hat{\sigma}_{i,kl})$

can be approximated by

$$\mathrm{E}(\hat{\sigma}_{i,kl}) \approx \mathrm{E}[(\hat{Z}_{i,k} - \nu_k)(\hat{Z}_{i,l} - \nu_l)] \approx \sigma_{kl} + \frac{C_{kl}}{M_i},$$

where $C_{kl} = \mathrm{E}(P_{1,K+1}^{-1}) - \frac{1}{2}\mathrm{Cov}(Z_{1,k}, P_{1,l}^{-1} - P_{1,K+1}^{-1}) - \frac{1}{2}\mathrm{Cov}(Z_{1,l}, P_{1,k}^{-1} - P_{1,K+1}^{-1})$.

References

Aitchison, J. (1986), *Coda: A Microcomputer Package for the Statistical Analysis Compositional Data*, Chapman and Hall.

Albert, R. and Barabási, A.-L. (2002), "Statistical mechanics of complex networks," *Reviews of modern physics*, 74, 47.

Arrigo, K. R. (2004), "Marine microorganisms and global nutrient cycles," *Nature*, 437, 349.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J.-M., et al. (2011), "Enterotypes of the human gut microbiome," *nature*, 473, 174.

Ban, Y., An, L., and Jiang, H. (2015), "Investigating microbial co-occurrence patterns based on metagenomic compositional data," *Bioinformatics*, 31, 3322–3329.

Billheimer, D., Guttorp, P., and Fagan, W. F. (2001), "Statistical interpretation of species composition," *Journal of the American statistical Association*, 96, 1205–1214.

Costea, P. I., Hildebrand, F., Arumugam, M., Bäckhed, F., Blaser, M. J., Bushman, F. D., De Vos, W. M., Ehrlich, S. D., Fraser, C. M., Hattori, M., et al. (2018), "Enterotypes in the landscape of gut microbial community composition," *Nature microbiology*, 3, 8.

Davin-Regli, A. et al. (2015), "Enterobacter aerogenes and Enterobacter cloacae; versatile bacterial pathogens confronting antibiotic treatment," *Frontiers in microbiology*, 6, 392.

Fang, H., Huang, C., Zhao, H., and Deng, M. (2015), "CCLasso: correlation inference for compositional data through Lasso," *Bioinformatics*, 31, 3172–3180.

Faust, K. and Raes, J. (2012), "Microbial interactions: from networks to models," *Nature Reviews Microbiology*, 10, 538.

Fioravanti, D., Giarratano, Y., Maggio, V., Agostinelli, C., Chierici, M., Jurman, G., and Furlanello, C. (2018), "Phylogenetic convolutional neural networks in metagenomics," *BMC bioinformatics*, 19, 49.

Friedman, J. and Alm, E. J. (2012), "Inferring correlation networks from genomic survey data," *PLoS computational biology*, 8, e1002687.

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, 9, 432–441.

Gorvitovskaia, A., Holmes, S. P., and Huse, S. M. (2016), "Interpreting Prevotella and Bacteroides as biomarkers of diet and lifestyle," *Microbiome*, 4, 15.

Jackson, D. A. (1997), "Compositional data in community ecology: the paradigm or peril of proportions?" *Ecology*, 78, 929–940.

Janda, J. M., Abbott, S. L., and McIver, C. J. (2016), "Plesiomonas shigelloides revisited," *Clinical microbiology reviews*, 29, 349–374.

Kamada, N., Chen, G. Y., Inohara, N., and Núñez, G. (2013), "Control of pathogens and pathobionts by the gut microbiota," *Nature immunology*, 14, 685.

Kohl, K. D., Weiss, R. B., Cox, J., Dale, C., and Denise Dearing, M. (2014), "Gut microbes of mammalian herbivores facilitate intake of plant toxins," *Ecology letters*, 17, 1238–1246.

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 computational biology*, 11, e1004226.

Lauritzen, S. L. (1996), *Graphical models*, vol. 17, Clarendon Press.

Mazmanian, S. K., Round, J. L., and Kasper, D. L. (2008), "A microbial symbiosis factor prevents intestinal inflammatory disease," *Nature*, 453, 620.

McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., Aksenov, A. A., Behsaz, B., Brennan, C., Chen, Y., et al. (2018), "American Gut: an Open Platform for Citizen Science Microbiome Research," *mSystems*, 3, e00031–18.

McMurdie, P. J. and Holmes, S. (2014), "Waste not, want not: why rarefying microbiome data is inadmissible," *PLoS computational biology*, 10, e1003531.

Roager, H. M., Licht, T. R., Poulsen, S. K., Larsen, T. M., and Bahl, M. I. (2014), "Microbial enterotypes, inferred by the prevotella-to-bacteroides ratio, remained stable during a 6-month randomized controlled diet intervention with the new nordic diet," *Appl. Environ. Microbiol.*, 80, 1142–1149.

Schneditz, G., Rentner, J., Roier, S., Pletz, J., Herzog, K. A., Bücker, R., Troeger, H., Schild, S., Weber, H., Breinbauer, R., et al. (2014), "Enterotoxicity of a nonribosomal peptide causes antibiotic-associated colitis," *Proceedings of the National Academy of Sciences*, 111, 13181–13186.

Xia, F., Chen, J., Fung, W. K., and Li, H. (2013), "A logistic normal multinomial regression model for microbiome compositional data analysis," *Biometrics*, 69, 1053–1063.

Yoon, G., Gaynanova, I., and Müller, C. L. (2019), "Microbial networks in SPRING-Semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data," *Frontiers in Genetics*, 10.

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

Yuan, M. and Lin, Y. (2007), "Model selection and estimation in the Gaussian graphical model," *Biometrika*, 94, 19–35.

Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012), "The huge package for high-dimensional undirected graph estimation in R," *Journal of Machine Learning Research*, 13, 1059–1062.

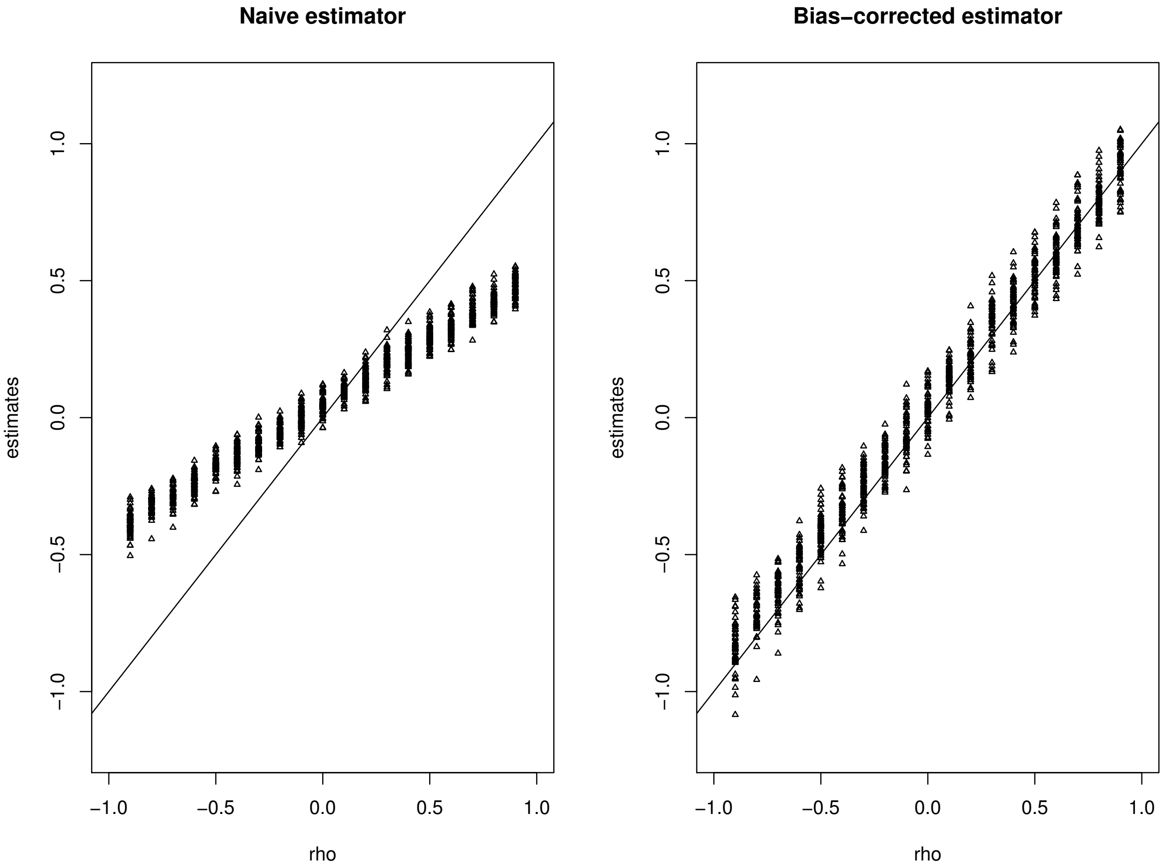

Fig. 1 Scatter plot of estimated and true covariances for the (1,2)-entry of the covariance matrix from a simulation study. Results are based on $K = 50$, $n = 500$, setting (M1) for the sequencing depth as described in the simulation setting, and a true correlation between the first and second taxa ranging from -0.9 to 0.9. Left panel: naive estimator; right panel: bias-corrected estimator.

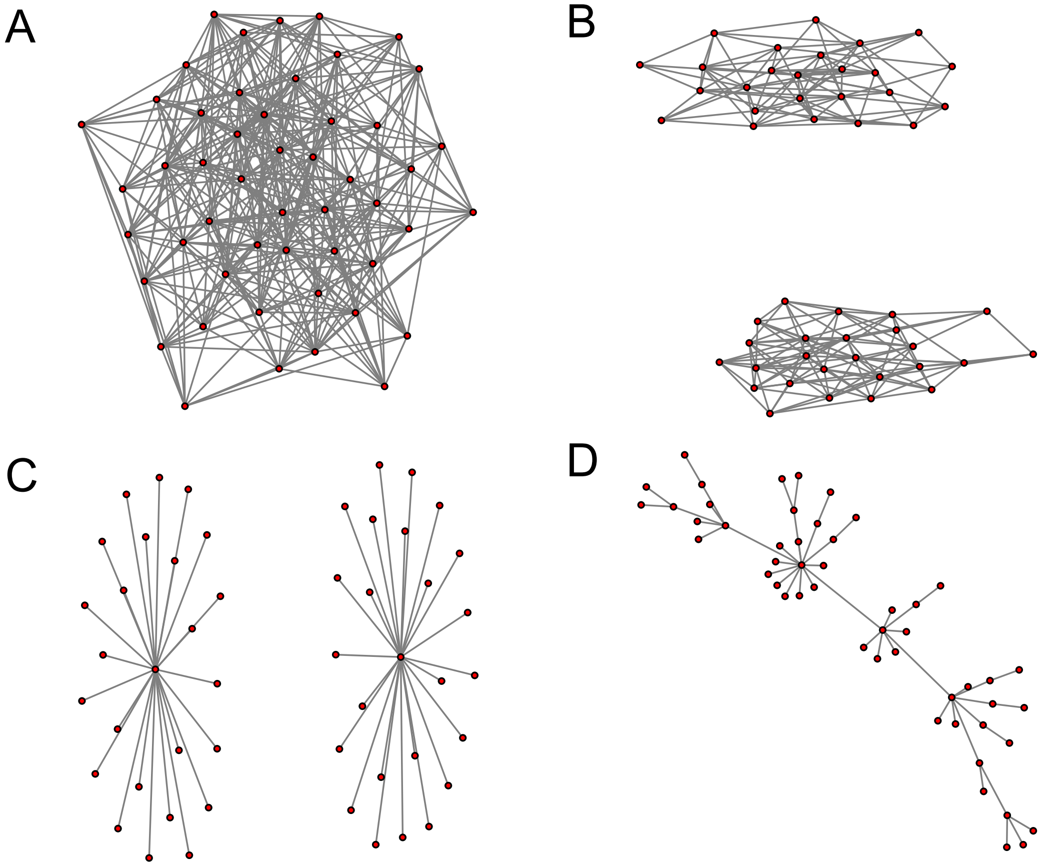

Fig. 2 The four network structures used in simulation studies. Panel A, random-edge network. Panel B, cluster network. Panel C, hub network; Panel D, scale-free network.

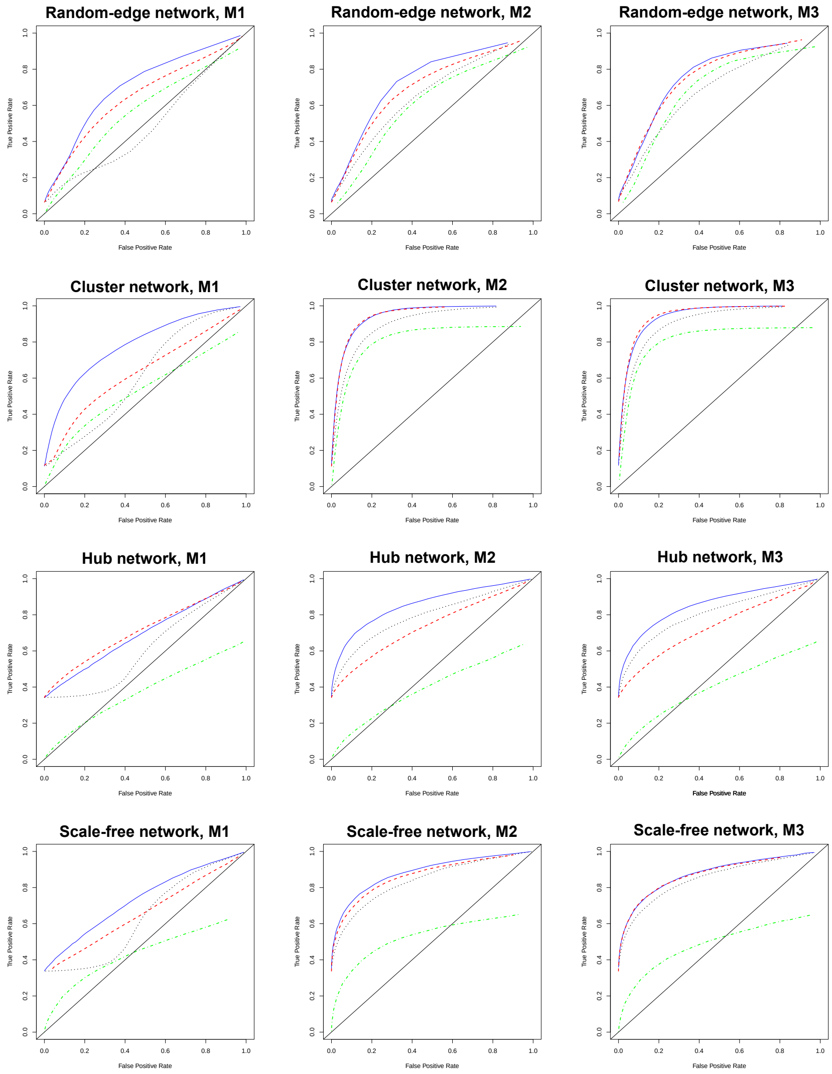

Fig. 3 BC-GLASSO enjoys more accurate network recovery than SPIEC-EASI in simulations. ROC curves for BC-GLASSO and SPIEC-EASI are compared for each simulation setting. Blue: BC-GLASSO. Black: SPIEC-EASI. Each row is a network structure. Each column is a configuration for the sequencing depth.

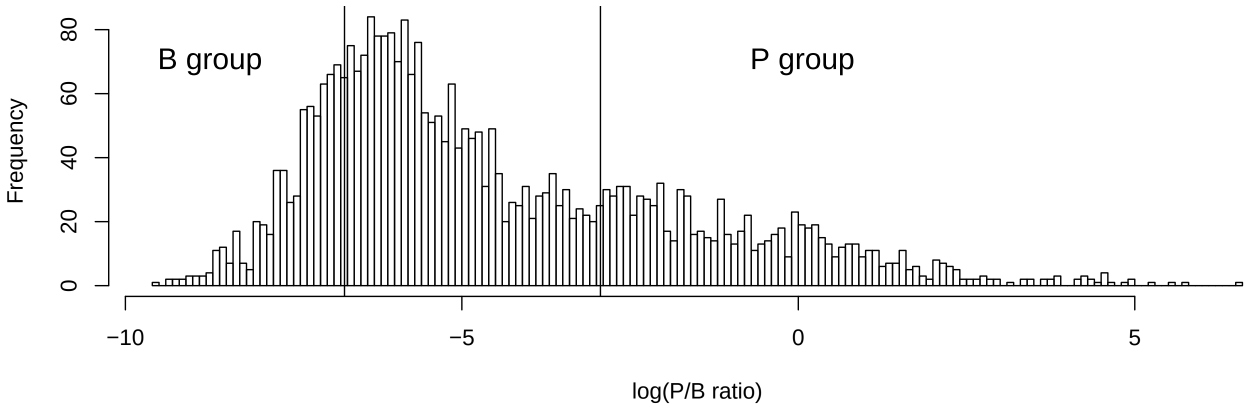

Fig. 4 Empirical distribution of P/B ratio from the AGP data and the regions in this distribution corresponding to the P group and the B group.

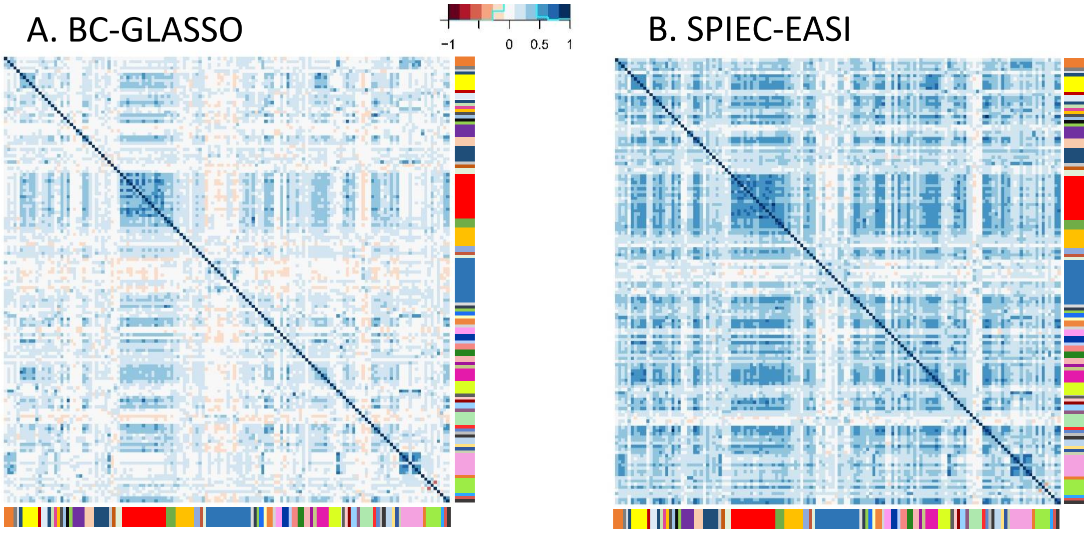

Fig. 5 Estimated correlation matrices for the P group of the AGP data based on BC-GLASSO and SPIEC-EASI. The correlation matrices are obtained based on the estimated covariance matrix used in the two methods. Left: BC-GLASSO. Right: SPIEC-EASI.

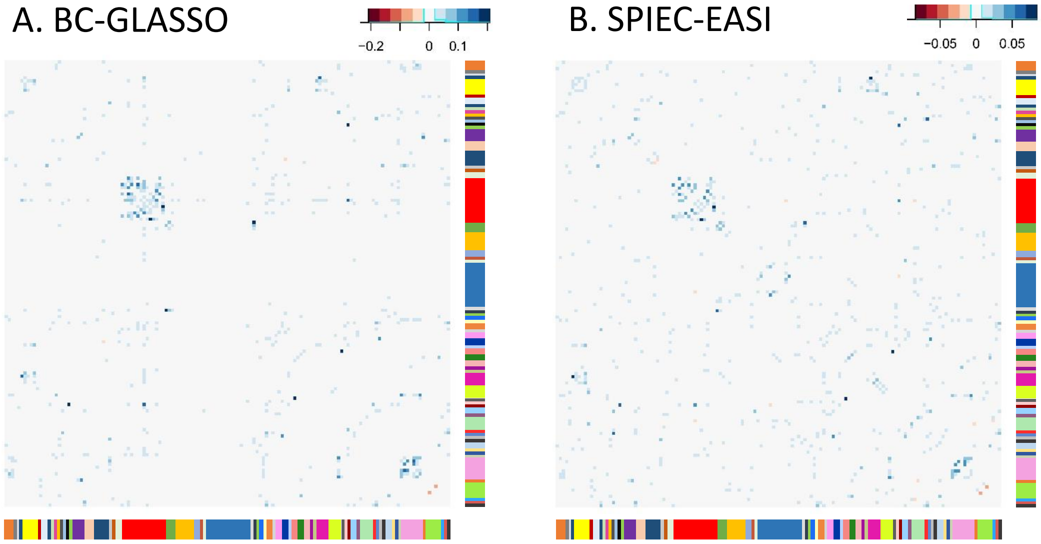

Fig. 6 Estimated inverse covariance matrices for the P group in the AGP data based on BC-GLASSO and SPIEC-EASI, on the negative scale. Diagonal entries are omitted. Left: BC-GLASSO. Right: SPIEC-EASI.

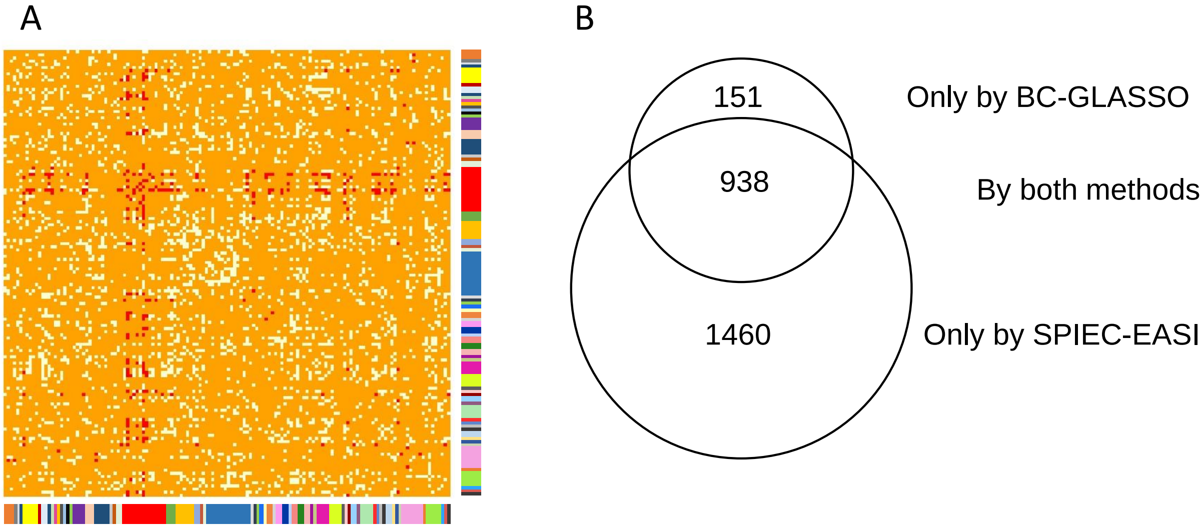

Fig. 7 Comparing the microbial networks based on BC-GLASSO and SPIEC-EASI for the P group of the AGP data. Panel A, network difference between two methods. Red - edges identified by BC-GLASSO but not by naive; White - edges identified by naive but not by BC-GLASSO; Orange - edges identified by either both or neither. Panel B, numbers of common and unique edges detected by BC-GLASSO and/or SPIEC-EASI.