JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 24, Number 7, 2017
© Mary Ann Liebert, Inc.
Pp. 699–708
DOI: 10.1089/cmb.2017.0054

gCoda: Conditional Dependence Network Inference for Compositional Data

HUAYING FANG,[1,2] CHENGCHENG HUANG,[3] HONGYU ZHAO,[4] and MINGHUA DENG[1,2,5]

ABSTRACT

The increasing quality and the reducing cost of high-throughput sequencing technologies for 16S rRNA gene profiling enable researchers to directly analyze microbe communities in natural environments. The direct interactions among microbial species of a given ecological system can help us understand the principles of community assembly and maintenance under various conditions. Compositionality and dimensionality of microbiome data are two main challenges for inferring the direct interaction network of microbes. In this article, we use the logistic normal distribution to model the background mechanism of microbiome data, which can appropriately deal with the compositional nature of the data. The direct interaction relationships are then modeled via the conditional dependence network under this logistic normal assumption. We then propose a novel penalized maximum likelihood method called gCoda to estimate the sparse structure of inverse covariance for latent normal variables to address the high dimensionality of the microbiome data. An effective Majorization-Minimization algorithm is proposed to solve the optimization problem in gCoda. Simulation studies show that gCoda outperforms existing methods (e.g., SPIEC-EASI) in edge recovery of inverse covariance for compositional data under a variety of scenarios. gCoda also performs better than SPIEC-EASI for inferring direct microbial interactions of mouse skin microbiome data.

Keywords: compositional data, direct interaction, inverse covariance matrix, microbial network, latent variable model, majorization-minimization algorithm.

1. INTRODUCTION

Microbes exist everywhere in natural environments; these microbiota can significantly impact the health of humans, and their interactions are implicated in varied human health conditions (Pflughoeft and Versalovic, 2012). Analysis of natural microbial communities can help us explore the way in which microbes affect their host or living environment. The high-throughput sequencing technologies, such as 16S rRNA gene profiling, provide an uncultivated microbial sampling strategy for diverse natural microbe

[1] LMAM, School of Mathematical Sciences, Peking University, Beijing, China.
[2] Center for Quantitative Biology, Peking University, Beijing, China.
[3] Institute of Urban Meteorology, China Meteorological Administration, Beijing, China.
[4] Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut.
[5] Center for Statistical Science, Peking University, Beijing, China.

communities (Kuczynski et al., 2012). The abundances of the underlying microbial species are quantified by operational taxonomic units (OTUs) counts. But the counts, which are usually converted into compositional data such as proportions based on total counts in one sample, only represent relative abundances of microbial species owing to different collection scales and various sequencing depths. This feature of microbiome data is called compositionality. Statistical analysis of such compositional data presents unique challenges since the constant sum's restriction can bring spurious results if it is ignored [e.g., correlation analysis (Pearson, 1897)]. In addition, the microbiome data are very high dimensional, with the number of measured OTUs often larger than the sample size. Such high dimensionality also presents statistical challenges for statistical inference, such as the inverse covariance estimation (Friedman et al., 2008).

An important goal of microbial ecology study is inferring the microbial interaction network in specific environments from the observed high-dimensional and compositional microbiome data (Faust and Raes, 2012; Weiss et al., 2016). Interactions can be divided into two types: direct interaction and indirect interaction. Direct interaction means the impact of one microbe on the other with no mediation through a third one, whereas indirect interaction is the impact between two microbes that are mediated or transmitted through a third one. Several methods have been proposed to infer the correlation network for microbiome studies (Faust et al., 2012; Friedman and Alm, 2012; Ban et al., 2015; Fang et al., 2015; Cao et al., 2016).

But compared with pairwise correlation dependences that include both direct and indirect interactions, researchers are often more concerned with the conditional dependences that describe the direct interactions (Friedman, 2004). Biswas et al. (2016) proposed an algorithm called MInt to learn direct interactions based on a Poisson-multivariate normal hierarchical model from microbiome sequencing experiments. But MInt does not explicitly account for the compositional nature of microbiome data. Kurtz et al. (2015) proposed an approximate method called SPIEC-EASI to infer direct interactions in microbiome studies. The key assumption of SPIEC-EASI is that the covariances of centered log-ratio transformations are near equal for absolute abundances and their compositional representations when the number of microbes is large enough. But this approximate assumption depends strongly on the condition number of the inverse covariance matrix. Recently, Yang et al. (2016) proposed a novel algorithm called mLDM to explore direct associations among microbes and between microbes and environmental factors from a hierarchical Bayesian model with sparsity constraints. However, mLDM lacks scalability and efficiency because numerous ancillary interim parameters are introduced in the complex hierarchical structure, which means that the computational burden rises considerably when the number of microbes becomes large.

In this article, we use a logistic normal distribution to model the generation mechanism of compositional data and propose a novel method called gCoda based on maximum likelihood with ℓ_1 penalty to estimate microbial conditional dependence structures that can describe direct interactions in microbial communities. One assumption of gCoda is that the latent absolute abundances follow a multivariate normal distribution in log scale. This assumption turns the conditional dependence inference problem into estimating the structure of the inverse covariance matrix. The other assumption is that the underlying ecological network is sparse, which can offset the information loss from both the constant sum's restriction of compositional data and the dimensionality problem of microbiome studies. We propose an effective Majorization-Minimization (MM) algorithm to solve the optimization problem involved in gCoda. The performance of gCoda is compared with SPIEC-EASI under various simulation scenarios. Simulation studies show that gCoda gives much better edge recovery than SPIEC-EASI for conditional dependence structures of compositional data. We also compare the inferred interaction networks between gCoda and SPIEC-EASI through a real microbiome data of mouse skin (Srinivas et al., 2013). The results of shuffled data show that the false positive count of gCoda is less than SPIEC-EASI. The gCoda is broadly applicable in many contexts when the observed data are compositional, and it's freely available from (see Reference 1) under LGPL v3.

2. METHODS

2.1. Logistic normal distribution for compositional data

Suppose the absolute abundance $y=(y_1, \ldots, y_p)^T$ of p species in a microbial community is modeled as a random vector, which cannot be directly observed in practice. Instead, only y's relative representation $x=(x_1, \ldots, x_p)^T$,

$$x_i = \frac{y_i}{\sum_{k=1}^{p} y_k}, \quad i = 1, \ldots, p, \tag{1}$$

is observed from biological experiments (Fang et al., 2015). The latent variable model in Equation (1) assumes that an unobserved total absolute abundance $w = \sum_{k=1}^{p} y_k$ exists, and it can be used to rebuild the absolute abundance from its observed compositional representation. Analysis of the absolute abundance y, rather than its compositional representation x, can overcome the constant sum's restriction $\sum_{k=1}^{p} x_k = 1$ that presents great challenges for correlation analysis (Pearson, 1897). The log-transformed data $\ln y = (\ln y_p, \ldots, \ln y_p)^T$ has linear relationships with $\ln x = (\ln x_p, \ldots, \ln x_p)^T$ from Equation (1),

$$\ln x = \ln y - \mathbf{1}_p \ln w, \tag{2}$$

where $\mathbf{1}_p$ is a $p \times 1$ vector of 1's. It is more convenient to deal with the log scale $\ln y$ than the original y because of the simple linear relationship in Equation (2). Another reason is that y should be positive whereas $\ln y$ does not have this restriction. So $\ln y$ is referred to as a latent variable in this article, and our goal is to infer the relationships among microbes from observed compositional data.

The random compositional vector x follows logistic normal distribution (Aitchison and Shen, 1980) if $\ln y$ follows a multivariate normal distribution $\mathcal{N}_p(\mu, \Sigma)$ with mean μ and nonsingular covariance matrix Σ. Under this logistical normal model, the structure of the inverse covariance matrix $\Omega = \Sigma^{-1}$ represents conditional dependence relationships among the elements of $\ln y$ since a zero entry $\Omega_{ij} = 0$ indicates that $\ln y_i$ and $\ln y_j$ are conditional independent given other left variables. The conditional dependence structure can describe direct interactions among microbial specials (Friedman, 2004). So inferring Ω from observed compositional data can help explore the direct interaction networks in microbiome studies.

2.2. gCoda

gCoda assumes that observed compositional data follow the logistic normal distribution and the direct interaction network of microbes is sparse. The first assumption, which can turn the inference of the direct interaction network of microbes into that of the structure of the inverse covariance of normal distribution, is about the distribution of compositional data. The second assumption, which can solve the under-determinated problem caused by compositionality (Fang et al., 2015) or dimensionality (Friedman et al., 2008), is about the edge density. Compared with absolute data, the totality information is lost for compositional data. So we cannot construct one unique inverse covariance from the observed compositional data without any constraint. If the true underlying inverse matrix is also sparse enough, we can try to find the most sparse one for the inverse covariances that all of them are corresponding to the observed compositional data. Since most microbial pairs are not expected to interact with each other directly when the number of microbes is large, the sparse assumption is reasonable in microbiome studies.

From $\ln y \sim \mathcal{N}_p(\mu, \Omega^{-1})$, the joint distribution of $(\ln w, x)$ is as follows:

$$f(\ln w, x) = (2\pi)^{-\frac{p}{2}} |\Omega|^{\frac{1}{2}} \prod_{i=1}^{p} x_i^{-1} \exp\left(-\frac{1}{2} Q\right),$$

where $Q = (\ln x + \mathbf{1}_p \ln w - \mu)^T \Omega (\ln x + \mathbf{1}_p \ln w - \mu)$ and $|\cdot|$ is the determinant of a matrix. For the sake of argument, the symbol x denotes the random variables $(x_1, x_2, \ldots, x_{p-1})^T$ when x appears on the left of a distribution function's expression and $x = (x_1, x_2, \ldots, x_p)^T$ when it appears on the right. So the conditional distribution of $\ln w$ given x is a one-dimensional normal distribution with mean $\frac{1}{\mathbf{1}_p^T \Omega \mathbf{1}_p} \mathbf{1}_p^T \Omega (\mu - \ln x)$ and variance $\frac{1}{\mathbf{1}_p^T \Omega \mathbf{1}_p}$. Let $F_0 = E_p - \frac{1}{p} \mathbf{1}_p \mathbf{1}_p^T$; then, the distribution of x can be got after integrating $f(\ln w, x)$ with respect to (w.r.t) $\ln w$,

$$f(x) = (2\pi)^{-\frac{p-1}{2}} \left(\frac{|\Omega|}{\mathbf{1}_p^T \Omega \mathbf{1}_p}\right)^{\frac{1}{2}} \prod_{i=1}^{p} x_i^{-1} \exp\left(-\tfrac{1}{2} Q_1\right),$$

where $Q_1 = (F_0 \ln x - F_0 \mu)^T \left(\Omega - \frac{\Omega \mathbf{1}_p \mathbf{1}_p^T \Omega}{\mathbf{1}_p^T \Omega \mathbf{1}_p}\right) (F_0 \ln x - F_0 \mu)$.

The negative log likelihood for (μ, Ω) based on the independent and identically distributed random samples $\{x^1, \ldots, x^n\}$ of the logistic normal distribution is as follows:

$$\mathcal{L}(\mu, \Omega) = -\ln \frac{|\Omega|}{\mathbf{1}_p^T \Omega \mathbf{1}_p} + \mathrm{tr}\left(S_0 \left(\Omega - \frac{\Omega \mathbf{1}_p \mathbf{1}_p^T \Omega}{\mathbf{1}_p^T \Omega \mathbf{1}_p} \right) \right),$$

up to a constant not depending on (μ, Ω), where $\mathrm{tr}(\cdot)$ is the trace of matrix, $S_0 = \frac{1}{n}\sum_{k=1}^{n} (F_0 \ln x^k - F_0 \mu)^{\otimes 2}$ and $a^{\otimes 2} = aa^T$ for a column vector a. Because we are more concerned about the estimation of Ω than μ, the sample mean of $F \ln x$ can be used as the estimation of $F\mu$ and we can get the negative log likelihood for Ω,

$$\mathcal{L}(\Omega) = -\ln|\Omega| + \ln(\mathbf{1}_p^T \Omega \mathbf{1}_p) + \mathrm{tr}\left(S \left(\Omega - \frac{\Omega \mathbf{1}_p \mathbf{1}_p^T \Omega}{\mathbf{1}_p^T \Omega \mathbf{1}_p} \right) \right),$$

where $S = \frac{1}{n}\sum_{k=1}^{n} (\ln x^k - \hat{\mu})^{\otimes 2}$ and $\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} \ln x^k$. Although the negative log likelihood for compositional data is derived from parametric distribution, this function $\mathcal{L}(\Omega)$ can be seen as the loss between observed compositional data and the inverse covariance in nonparametric situations.

It has been pointed out that the estimation problem of the latent variable model for compositional data is unidentifiable if there are no more assumptions about the unknown parameters (Fang et al., 2015). In addition, the under-determined problem also arises if the sample size is smaller than the dimension of variables (Friedman et al., 2008). Here, we assume that only few edges exist in the conditional dependence network, that is, Ω is sparse. A commonly used approach of sparse structures is to add ℓ_1 penalty to some loss function that measures the fitting of the observed data (Tibshirani, 1996). So, we consider the following objective function combining negative log likelihood and ℓ_1 penalty,

$$f(\Omega) = \mathcal{L}(\Omega) + \lambda_n \|\Omega\|_1,$$

where $\|\Omega\|_1 = \sum_{i=1}^{p}\sum_{j=1}^{p} \Omega_{ij}$ and the tuning parameters $\lambda_n > 0$ are used to balance the model fitting of observed data and the sparse degree of Ω. Then, gCoda aims at finding the maximum likelihood estimation with sparse ℓ_1 penalty as follows:

$$\hat{\Omega} = \underset{\Omega \succ 0}{\arg\min}\, f(\Omega) = \underset{\Omega \succ 0}{\arg\min}\, \mathcal{L}(\Omega) + \lambda_n \|\Omega\|_1, \tag{3}$$

where $\Omega \succ 0$ means that Ω should be positive definite. Since the negative log likelihood function $\mathcal{L}(\Omega)$ is not convex, the optimization problem involved in Equation (3) is not convex when λ_n is small. Thus, only a local minimization can be got as the estimation of inverse covariance. The following algorithm for gCoda always provides an approximate estimation for Ω in practice.

2.3. *MM algorithm and choice of λ_n*

The optimization problem in Equation (3) is difficult because the objective function $f(\Omega)$ is neither convex nor smooth, and the solution requires being positive definite. Here, an efficient MM algorithm is developed to solve the constrained optimization problem in gCoda. The MM algorithm guarantees that the objective function decreases in each step until a local optimum or a saddle point is reached by minimizing a series of surrogate functions when optimizing surrogate functions is much easier than direct optimization for the objective function. At the kth step of the MM algorithm, $g(\theta|\theta_k)$ is called a majorizing function of $f(\theta)$ at θ_k if $g(\theta|\theta_k) \geq f(\theta), \forall\theta$ and $g(\theta_k|\theta_k) = f(\theta_k)$. The MM algorithm updates θ via $\theta_{k+1} = \arg\min_\theta g(\theta|\theta_k)$. This iterative procedure guarantees that $f(\theta_k)$ decreases in each iteration (Lange et al., 2000). We construct the following majorizing function for $f(\Omega)$ in gCoda,

$$\begin{aligned} g(\Omega|\Omega_k) = &-\ln|\Omega| + \mathrm{tr}\left(\Omega \left(E_p - \frac{\mathbf{1}_p \mathbf{1}_p^T \Omega_k}{\mathbf{1}_p^T \Omega_k \mathbf{1}_p} \right) S \left(E_p - \frac{\Omega_k \mathbf{1}_p \mathbf{1}_p^T}{\mathbf{1}_p^T \Omega_k \mathbf{1}_p} \right) \right) \\ &+ \ln(\mathbf{1}_p^T \Omega_k \mathbf{1}_p) + \frac{1}{\mathbf{1}_p^T \Omega_k \mathbf{1}_p} (\mathbf{1}_p^T \Omega \mathbf{1}_p - \mathbf{1}_p^T \Omega_k \mathbf{1}_p) + \lambda_n \|\Omega\|_1. \end{aligned}$$

It is obvious that $g(\Omega_k|\Omega_k) = f(\Omega_k)$. From the concavity of the logarithm function and Cauchy–Schwarz inequality, we can get $g(\Omega|\Omega_k) \geq f(\Omega)$. So, $g(\Omega|\Omega_k)$ is one majorizing function for $f(\Omega)$ at Ω_k. And minimizing $g(\Omega|\Omega_k)$ w.r.t Ω is a standard graphical lasso problem since

$$\Omega_{k+1} = \underset{\Omega \succ 0}{\arg\min}\ g(\Omega|\Omega_k) = \underset{\Omega \succ 0}{\arg\min} - \ln|\Omega| + \mathrm{tr}(\Omega S_k) + \lambda_n \ \|\Omega\|_1,$$

where $S_k = \left(E_p - \frac{\mathbf{1}_p \mathbf{1}_p^T \Omega_k}{\mathbf{1}_p^T \Omega_k \mathbf{1}_p}\right) S \left(E_p - \frac{\Omega_k \mathbf{1}_p \mathbf{1}_p^T}{\mathbf{1}_p^T \Omega_k \mathbf{1}_p}\right) + \frac{1}{\mathbf{1}_p^T \Omega_k \mathbf{1}_p} \mathbf{1}_p \mathbf{1}_p^T$. So, the MM algorithm decomposes the optimization problem (3) into a series of graphical lasso problems that can be solved effectively via the blockwise coordinate descent approach (Friedman et al., 2008). The following algorithm summarizes details to carry out the MM algorithm for gCoda mentioned earlier.

```
(1). Initialize $\Omega_0$ and set $k \leftarrow 0$.
(2). Repeat (a)–(c) until $\Omega_k$ converges:
    (a). Compute $S_k$;
    (b). Solve $\Omega_{k+1} = \arg\min_{\Omega \succ 0} - \ln|\Omega| + \mathrm{tr}(\Omega S_k) + \lambda_n \ \|\Omega\|_1$ via glasso algorithm;
    (c). $k \leftarrow k+1$.
(3). Return converged $\Omega_k$ as $\hat{\Omega}$ defined in Equation (3).
```

The positive parameter λ_n in Equation (3) controls the balance between the likelihood of observed data and the sparsity of inverse covariance. Here, λ_n is selected via extended Bayesian information criteria (EBIC, Chen and Chen, 2008). First, for given λ_n, compute $\hat{\Omega}(\lambda_n)$ in Equation (3) and the EBIC score $\mathrm{EBIC}_{0.5}(\lambda_n) = n\mathcal{L}(\hat{\Omega}(\lambda_n)) + \#\{\hat{\Omega}(\lambda_n)\}(\ln n + 2\ln p)$, where $\#\{\hat{\Omega}(\lambda_n)\}$ is the number of edges in the network represented by $\hat{\Omega}(\lambda_n)$. Then, $\lambda^* = \arg\min_{\lambda_n} \mathrm{EBIC}_{0.5}(\lambda_n)$ is chosen for gCoda.

3. RESULTS

3.1. Simulation studies

The performance of gCoda and SPIEC-EASI is compared via compositional data rather than counts data in simulation studies since they have the same assumption in Equation (1). Two variants of SPIEC-EASI are denoted as SE(MB) and SE(GL) that infer interaction networks via neighborhood and covariance selection, respectively. The area under the receiver operating characteristic (ROC) curve (AUC) is used to assess the performance of gCoda and SPIEC-EASI in recovering nonzero entries in the sparse inverse covariance.

The compositional data are generated from a logistic normal distribution with given mean μ and inverse covariance Ω as $\ln y \sim \mathcal{N}_p(\mu, \Omega^{-1})$ and $x_i = y_i / \sum_{i=1}^{k} y_i$, $1 \leq i \leq p$. The mean μ controls the unbalance of components and is generated from a uniform distribution in $[-0.5, 0.5]^p$. The following six common sparse network structures for Ω are used in our simulations:

1. *Random graph*: Two nodes are connected with probability 0.2 and strength ±0.15 under equal possibility.
2. *Neighbor graph*: Randomly select p points in the [0, 1] plane and connect 10 nearest neighbors for each point, with strength ±0.8 being equally probable.
3. *Band graph*: Connect pair (i, j) if $|i-j| \leq 4$ and the strength is set as $-0.8, 0.6, 0.4, -0.2$ when $|i-j| = 1, 2, 3, 4$.
4. *Hub graph*: Randomly select three points as hubs. The hubs are connected to nonhubs with probability 0.8 and strength 0.2, whereas pairs in nonhubs are connected with probability 0.2 and strength 0.2.
5. *Block graph*: Split p nodes into five blocks equally. Pairs in the same blocks are connected with probability 0.3 and strength 0.5, whereas connecting pairs in different blocks are connected with probability 0.1 and strength 0.25.
6. *Scale-free graph*: The B-A algorithm (Barabási and Albert, 1999) is used to build a scale-free network. Start with a single node and then add $p-1$ nodes one by one. In each step, the new node is connected with three randomly selected old nodes, with probability-related nodes' degrees in the current graph. The strength is generated from a uniform distribution in $[-0.8, -0.6] \cup [0.6, 0.8]$.

The diagonal elements of Ω are set large enough to make Ω positive definite and normalized all as 1. The number of nodes p is set at 50, whereas the sample size is varied ($n = 100$, 200, and 500). For each combination of graph structure and sample size, we repeat simulations 20 times and calculate the averaged AUC values as the final performance of gCoda and SPIEC-EASI.

Table 1 summarizes AUC values of different graph structures and sample sizes for gCoda and SPIEC-EASI. For each simulation setting, AUC of gCoda is larger than both variants of SPIEC-EASI, that is, gCoda outperforms SPIEC-EASI in the edge recovery of interaction networks. For each combination of graph structure and method, AUC increases as the sample size increases, except for the scale-free network and SE(MB). We can find that the performance of SPIEC-EASI is dependent on the network structure. SPIEC-EASI works well for random, neighbor, band, and block graphs, but bad for hub and scale-free graphs. Our gCoda is more robust than SPIEC-EASI from the simulation results. More detailed results for ROC curves are shown in Figure 1. For small sample size and some specific graph structures, including random, neighbor and band, ROC curves for gCoda and SPIEC-EASI are very close. The difference of performances between gCoda and SPIEC-EASI increases as the sample size increases. The difference for hub, block, and scale-free graphs is larger than the other three graph structures. SE(MB) performs worse than random guess in the beginning part of ROC curves for the hub graph. The performance of SPIEC-EASI is unstable for different graph structures.

We also consider the situation when the number of microbes is greater than or equal to the sample size from simulation studies (Supplementary Fig. S1). The performances of gCoda and SPIEC-EASI are similar in most graphs, whereas gCoda outperforms SPIEC-EASI for the scale-free graph. We also explore the effect of compositionality on the estimation of the inverse covariance matrix for observed compositional data through simulation studies (Supplementary Fig. S2). The results suggest that we cannot directly treat the compositional data as absolute abundances when inferring the interaction network from observed data.

3.2. *Analysis of mouse skin microbiome data*

We apply gCoda to infer the direct interaction networks of microbes for a mouse skin microbiome data from a study population of 261 mice (Srinivas et al., 2013). According to health conditions of the skin's immunizations, the samples in this data are divided into three groups: 78 nonimmunized controls (Control), 119 immunized healthy individuals (Healthy), and 64 immunized epidermolysis bullosa acquisita (EBA) individuals. The data are further filtered by removing OTUs that are represented in less than 60% samples and removing samples for which >60% OTUs are 0s. A total of 229 samples and 60 OTUs remain after data filtering. We add all OTU counts by the maximum rounding error 0.5 and then normalize the counts into compositional data. The parameters of stable selections for SPIEC-EASI are set according to the examples in https://github.com/zdk123/SpiecEasi. Both gCoda and SPIEC-EASI use their default cut-off values to get the final inferred networks. The numbers of inferred edges by these algorithms are comparable for the Control and EBA groups, whereas the inferred network of microbes by gCoda is denser than SPIEC-EASI for the Healthy group.

Table 1. Performance Comparisons of gCoda and SPIEC-EASI via Area-under-the-curve Values in Simulation Studies

		Network structure					
N	*Method*	*Random*	*Neighbor*	*Band*	*Hub*	*Block*	*Scale free*
100	gCoda	0.714	0.795	0.848	0.696	0.803	0.752
	SE(MB)	0.708	0.760	0.833	0.526	0.745	0.605
	SE(GL)	0.712	0.770	0.834	0.543	0.692	0.611
300	gCoda	0.800	0.893	0.900	0.793	0.849	0.781
	SE(MB)	0.730	0.860	0.865	0.579	0.786	0.651
	SE(GL)	0.755	0.842	0.864	0.588	0.712	0.634
500	gCoda	0.837	0.912	0.920	0.822	0.857	0.789
	SE(MB)	0.732	0.889	0.874	0.594	0.798	0.648
	SE(GL)	0.774	0.863	0.875	0.604	0.718	0.639

The area under the curve value is the area under the receiver operating characteristic curve. SE(MB) and SE(GL) are two variants of SPIEC-EASI. The results are the averages over 20 simulation runs.

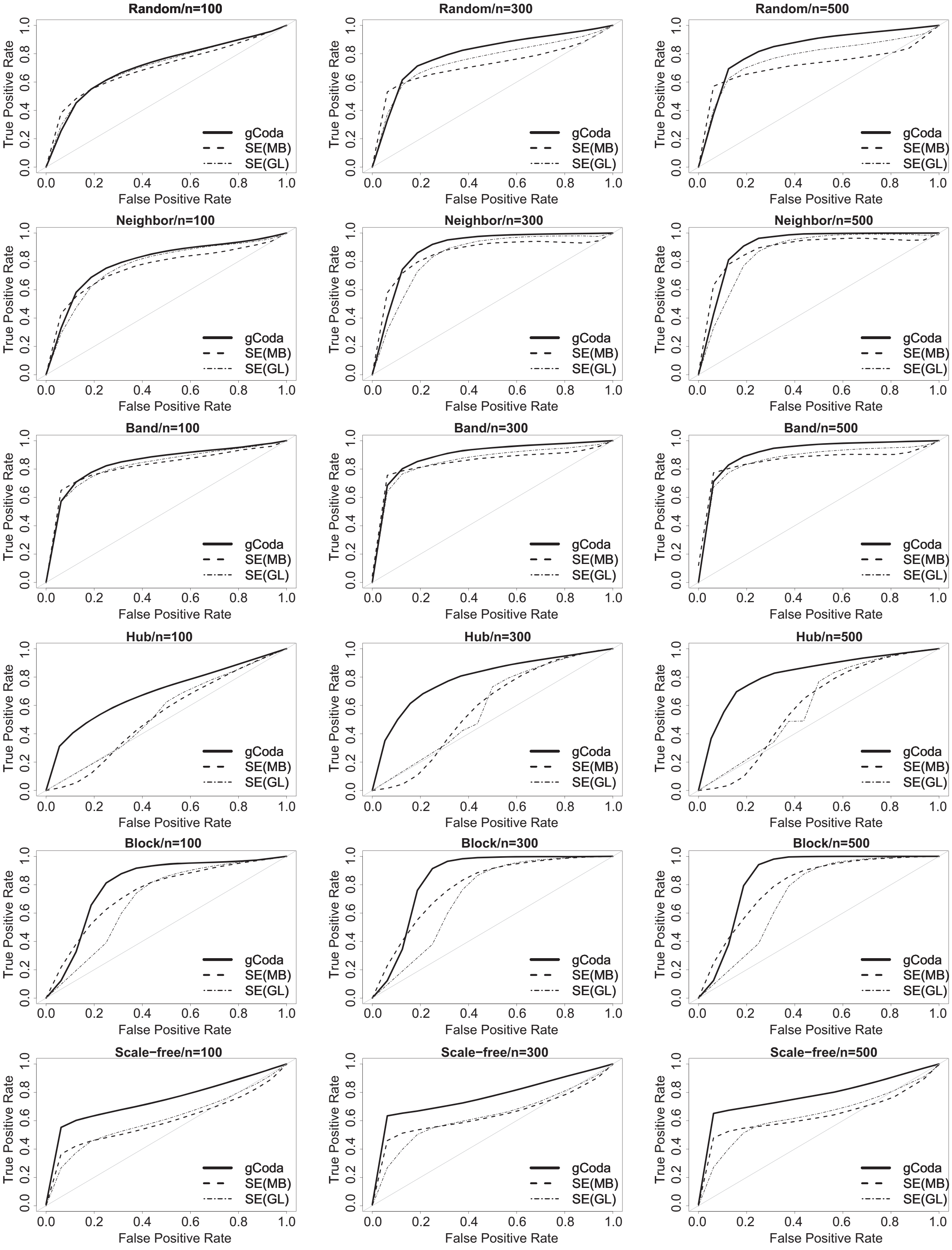

FIG. 1. ROC curves of gCoda and SPIEC-EASI. Each row corresponds to a specific graph structure, whereas each column corresponds to a specific sample size. SE(MB) and SE(GL) are two variants of SPIEC-EASI. These results are averaged over 20 replications with the same simulation setting. ROC, receiver operating characteristic.

Table 2. Performance Comparisons of gCoda and SPIEC-EASI via False Positive Count and Running Time(s) in the Mouse Skin Data

	False positive count			*Time(s)*		
Method	*Control*	*Healthy*	*EBA*	*Control*	*Healthy*	*EBA*
gCoda	3.2	1.8	2.5	1.29	0.35	3.63
SE(MB)	5.8	7.1	9.4	67.41	66.52	64.52
SE(GL)	5.6	7.1	9.3	74.10	68.24	76.91

The false positive count is the false positive edge's count for the shuffled OTU count matrix. Time(s) is the running time for shuffled data. SE(MB) and SE(GL) are two variants of SPIEC-EASI. The results are the averages over 10 replications.

Since no prior information of true taxon-taxon interaction networks exists in real data, we compare gCoda and SPIEC-EASI from the following aspects. The first is the false positive count of shuffled OTU tables. It is supposed to find no interaction among species from shuffled data, so the count of edges inferred by gCoda or SPIEC-EASI can measure the false positive count in real data. The second is the running time of gCoda and SPIEC-EASI for shuffled data under a Linux workstation: Intel(R) Xeon(R) E5640 (2.66 GHz) CPU and 16 GB MEM. All of these measures are replicated 10 times, and the averaged results are summarized in Table 2. We can find that the false positive count of gCoda is less than SPIEC-EASI. And gCoda is much faster than SPIEC-EASI since the stable selection procedure of SPIEC-EASI is time-consuming.

We also compare the overlaps among networks for different groups and different methods. Figure 2 summarizes the results of shared edges between different networks under various situations via Venn diagrams. The number of edges shared among three mouse groups for gCoda is larger than the two variants of SPIEC-EASI. The network size for the Healthy group via gCoda is much larger than for SPIEC-EASI. One possible reason of this phenomenon is that the Healthy group is in the middle of the other two groups,

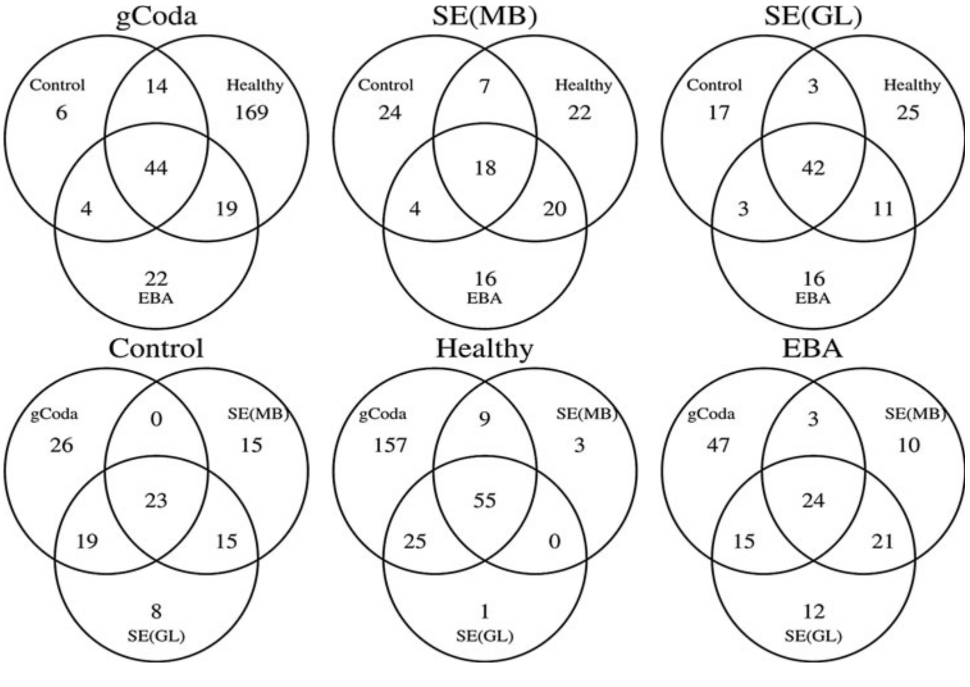

FIG. 2. Venn diagrams of shared edges among the inferred networks by gCoda and SPIEC-EASI. The first row represents overlaps for the Control, Healthy, and EBA groups that are inferred from the same algorithm, whereas the second row represents overlaps of three algorithms for the same group. SE(MB) and SE(GL) are two variants of SPIEC-EASI.

and many connections among microbes are needed to maintain the intermediary role. Since the two variants of SPIEC-EASI are based on the same approximation formula, the overlap between these two variants is large for all of the three mouse groups.

4. DISCUSSION

High-throughput sequencing technologies provide unprecedented opportunities to explore the relationships among microbes in natural environments. But inferring direct taxon-taxon interaction networks from sequencing data is still difficult since only relative abundances of microbes can be observed from microbome studies and the sample size is often smaller than the number of microbes. Here, we propose a novel method called gCoda to infer the sparse direct interaction network among microbes from the logistic normal distribution of observed compositional data. From simulations with various graph structures and analysis of real microbiome data, gCoda is found to be more accurate and robust in edge recovery than existing methods (e.g., SPIEC-EASI).

Our gCoda is derived from the penalized maximum likelihood and can lead to a sparse inverse covariance matrix to construct the direct interaction network. It is more stable and accurate with less computation time than existing methods, such as SPIEC-EASI. We believe that gCoda will be broadly applicable in many other contexts where compositional data are observed.

ACKNOWLEDGMENTS

This work was supported by the National Key Basic Research Project of China (No. 2015CB910303), the National Key Research and Development Program of China (No. 2016YFA0502300), and the National Natural Science Foundation of China (Nos. 31171262, 31428012, 31471246). H.F.'s work was supported, in part, by a fellowship from the Graduate School of Peking University. H.Z.'s work was supported, in part, by NIH GM59507.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

https://github.com/huayingfang/gCoda

Aitchison, J., and Shen, S.M. 1980. Logistic-normal distributions: Some properties and uses. *Biometrika* 67, 261–272.

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

Barabási, A.-L., and Albert, R. 1999. Emergence of scaling in random networks. *Science* 286, 509–512.

Biswas, S., McDonald, M., Lundberg, D.S., et al. 2016. Learning microbial interaction networks from metagenomic count data. *J Comput Biol* 23, 526–535.

Cao, Y., Lin, W., and Li, H. 2016. Large covariance estimation for compositional data via composition-adjusted thresholding. *arXiv preprint arXiv:1601.04397*.

Chen, J., and Chen, Z. 2008. Extended bayesian information criteria for model selection with large model spaces. *Biometrika* 95, 759–771.

Fang, H., Huang, C., Zhao, H., et al. 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. *Nat. Rev. Microbiol.* 10, 538–550.

Faust, K., Sathirapongsasuti, J.F., Izard, J., et al. 2012. Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.* 8, e1002606.

Friedman, J., and Alm, E.J. 2012. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* 8, e1002687.

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

Friedman, N. 2004. Inferring cellular networks using probabilistic graphical models. *Science* 303, 799–805.

Kuczynski, J., Lauber, C.L., Walters, W.A., et al. 2012. Experimental and analytical tools for studying the human microbiome. *Nat. Rev. Genet.* 13, 47–58.

Kurtz, Z.D., Müller, C.L., Miraldi, E.R., et al. 2015. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* 11, e1004226.

Lange, K., Hunter, D.R., and Yang, I. 2000. Optimization transfer using surrogate objective functions. *J. Comput. Graph. Stat.* 9, 1–20.

Pearson, K. 1897. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proc. Roy. Soc. Lond.* 60, 489–502.

Pflughoeft, K.J., and Versalovic, J. 2012. Human microbiome in health and disease. *Annu. Rev. Pathol. Mech. Dis.* 7, 99–122.

Srinivas, G., Möller, S., Wang, J., et al. 2013. Genome-wide mapping of gene–microbiota interactions in susceptibility to autoimmune skin blistering. *Nat. Commun.* 4, 2462.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. B Met.* 58, 267–288.

Weiss, S., Van Treuren, W., Lozupone, C., et al. 2016. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.* 10, 1669–1681.

Yang, Y., Chen, N., and Chen, T. 2016. mLDM: A new hierarchical bayesian statistical model for sparse microbioal association discovery. *bioRxiv* page 042630.

Address correspondence to:
Prof. Minghua Deng
LMAM School of Mathematical Sciences
Peking University
Beijing 100871
China

E-mail: dengmh@pku.edu.cn