Microbial Network Recovery by Compositional Graphical Lasso

Chuan Tian, Duo Jiang, Thomas Sharpton, Yuan Jiang

May 28, 2020

Abstract

Network models such as graphical models have become a useful approach to studying the interactions between microbial taxa given the microbiome data deluge. Recently, 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 presented by error-prone counts. Adding to the challenge is that the sum of the counts within each sample, termed "sequencing depth", can vary drastically across samples. To address these issues, we adopt a logistic normal multinomial model explicitly incorporating the sequencing depth and develop an algorithm iterated between Newton-Raphson and graphical lasso for model estimation. We call this new approach "compositional graphical lasso". We have established the convergence of the algorithm. Additionally, we illustrate the advantage of compositional graphical lasso in comparison to current methods under a variety of simulation scenarios and also demonstrate the applicability of compositional graphical lasso to a real 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 (OTUs, 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. Third, the data are high-dimensional in nature. It is likely that the number of OTUs are far more than the number of samples in any biological experiment.

When such abundance 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 two 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 study 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; Banerjee et al., 2008; Friedman et al., 2008). However, both the counting and the compositional features of the microbiome abundance data have violated the multivariate normality assumption.

SPEIC-EASI was probably the first method that was proposed to estimate sparse microbial network based on conditional dependency (Kurtz et al., 2015). It first performs a central log-ratio transformation on the observed counts (Aitchison, 1986), and then apply graphical lasso to the transformed data. SPEIC-EASI avoided the violation of the distributional assumption by transforming counts into continuous log ratios but it did not address the compositional feature of the data. Recently, mLDM was proposed as a three-level hierarchical model (Yang et al., 2016), which hierarchically models the OTU counts by a multinomial distribution, the multinomial probabilities by a Dirichlet distribution, and the Dirichlet parameters by a lognormal distribution. While mLDM have taken into account the compositional count nature, its algorithm and resultant estimators lack theoretical justifications. Moreover, the model is very complex so its interpretability is rather limited.

In this paper, we adopt the logistic normal multinomial distribution to model the compositional count data (Aitchison, 1986; Billheimer et al., 2001; Xia et al., 2013). Compared to the three-level hierarchical model, this model only has two levels, thus is more interpretable. We further develop an algorithm iterated between Newton-Raphson and graphical lasso for model estimation. We call this new approach "compositional graphical lasso". We have further established the theoretical convergence of the algorithm, which was not available for either SPEIC-EASI or mLDM. Additionally, we illustrate the advantage of compositional graphical lasso in comparison to current methods under a variety of simulation scenarios and also demonstrate the applicability of compositional graphical lasso to a human microbiome data set.

2 Compositional Graphical Lasso

2.1 Logistic Normal Multinomial 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:

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

where $\mathbf{p}_i = (p_{i,1}, \ldots, p_{i,K+1})'$ are the multinomial probabilities for all taxa and $\sum_{k=1}^{K+1} p_{i,k} = 1$.

In addition, we choose one taxon, without loss of generality, the $(K+1)$-th taxon as a reference for all the others and then apply the additive log-ratio 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 $\mathbf{z}_i = (z_{i,1}, \ldots, z_{i,K})'$ for $i = 1, \ldots, n$, and further assume that they follow an i.i.d. multivariate normal distribution

$$\mathbf{z}_1, \ldots, \mathbf{z}_n \overset{iid}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{3}$$

where $\boldsymbol{\mu}$ is the mean, $\boldsymbol{\Sigma}$ is the covariance matrix, and $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is the inverse covariance matrix or the precision matrix.

The above model in (1)–(3) is often referred to as the logistic normal multinomial model. In this model, 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. Therefore, the logistic normal multinomial model is a hierarchical model with two levels.

The logistic normal multinomial model has a long history in modeling compositional count data and it has also been applied to analyze the microbiome abundance data. For example, Xia et al. (2013) proposed a penalized regression under this model to identify a subset of covariates that are associated with the taxon composition. Our objective is different from Xia et al. (2013) as we aim to reveal the microbial interaction network by finding a sparse estimator of the inverse covariance matrix $\boldsymbol{\Omega}$ in (3). It is also noteworthy that Jiang et al. (2020) has the same objective as ours. However, Jiang et al. (2020) did not make full use of the logistic normal multinomial model as it focused on correcting the bias of a naive estimator of the $\boldsymbol{\Sigma}$ that does not require the logistic normal assumption. By contrast, we aim to find an estimator of $\boldsymbol{\Omega}$ directly based on the logistic normal multinomial model.

2.2 Objective Function

From the logistic normal multinomial model in (1)–(3), we can write the logarithm of the posterior distribution of $\mathbf{z}_i$ given the data $\mathbf{x}_i$, ignoring a term that only depends on $\mathbf{x}_i$,

$$\begin{aligned}
&\log[f_{\boldsymbol{\mu},\boldsymbol{\Omega}}(\mathbf{z}_i|\mathbf{x}_i)] \propto \log[f_{\boldsymbol{\mu},\boldsymbol{\Omega}}(\mathbf{x}_i,\mathbf{z}_i)] \\
\propto &\sum_{k=1}^{K+1} x_{i,k}\log p_{i,k} + \frac{1}{2}\log[\det(\boldsymbol{\Omega})] - \frac{1}{2}(\mathbf{z}_i-\boldsymbol{\mu})'\boldsymbol{\Omega}(\mathbf{z}_i-\boldsymbol{\mu}) \\
= &\sum_{k=1}^{K} x_{i,k}z_{i,k} - M_i\log\left(\sum_{k=1}^{K} e^{z_{i,k}}+1\right) + \frac{1}{2}\log[\det(\boldsymbol{\Omega})] - \frac{1}{2}(\mathbf{z}_i-\boldsymbol{\mu})'\boldsymbol{\Omega}(\mathbf{z}_i-\boldsymbol{\mu}).
\end{aligned}$$

By independence between all the samples, the logarithm of the posterior distribution of $(\mathbf{z}_1,\ldots,\mathbf{z}_n)$ given the data $(\mathbf{x}_1,\ldots,\mathbf{x}_n)$ can be written as

$$\begin{aligned}
&\log[f_{\boldsymbol{\mu},\boldsymbol{\Omega}}(\mathbf{z}_1,\ldots,\mathbf{z}_n|\mathbf{x}_1,\ldots,\mathbf{x}_n)] \propto \sum_{i=1}^{n}\log[f_{\boldsymbol{\mu},\boldsymbol{\Omega}}(\mathbf{x}_i,\mathbf{z}_i)] \\
\propto &\sum_{i=1}^{n}\left[\sum_{k=1}^{K} x_{i,k}z_{i,k} - M_i\log\left(\sum_{k=1}^{K} e^{z_{i,k}}+1\right)\right] + \frac{n}{2}\log[\det(\boldsymbol{\Omega})] - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{z}_i-\boldsymbol{\mu})'\boldsymbol{\Omega}(\mathbf{z}_i-\boldsymbol{\mu}).
\end{aligned}$$

Given the multivariate normal parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$, one can maximize the posterior distribution with respect to $(\mathbf{z}_1,\ldots,\mathbf{z}_n)$. This leads to the posterior mode $(\hat{\mathbf{z}}_1,\ldots,\hat{\mathbf{z}}_n)$ and

returns the logarithm of the maximum posterior distribution as

$$\log[f_{\boldsymbol{\mu},\boldsymbol{\Omega}}(\hat{\mathbf{z}}_1,\ldots,\hat{\mathbf{z}}_n|\mathbf{x}_1,\ldots,\mathbf{x}_n)] = \max_{\mathbf{z}_1,\ldots,\mathbf{z}_n} \log[f_{\boldsymbol{\mu},\boldsymbol{\Omega}}(\mathbf{z}_1,\ldots,\mathbf{z}_n|\mathbf{x}_1,\ldots,\mathbf{x}_n)].$$

Since our objective is to find a sparse estimator of the inverse covariance matrix $\boldsymbol{\Omega}$, we further maximizes the above maximum posterior distribution over $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ with a L_1 penalty on $\boldsymbol{\Omega}$, or equivalently,

$$\begin{aligned} &\min_{\boldsymbol{\mu},\boldsymbol{\Omega}} -\frac{1}{n}\log[f_{\boldsymbol{\mu},\boldsymbol{\Omega}}(\hat{\mathbf{z}}_1,\ldots,\hat{\mathbf{z}}_n|\mathbf{x}_1,\ldots,\mathbf{x}_n)] + \lambda\|\boldsymbol{\Omega}\|_1 \\ = &\min_{\boldsymbol{\mu},\boldsymbol{\Omega}} \min_{\mathbf{z}_1,\ldots,\mathbf{z}_n} -\frac{1}{n}\log[f_{\boldsymbol{\mu},\boldsymbol{\Omega}}(\mathbf{z}_1,\ldots,\mathbf{z}_n|\mathbf{x}_1,\ldots,\mathbf{x}_n)] + \lambda\|\boldsymbol{\Omega}\|_1. \end{aligned} \tag{4}$$

The above derivations suggest that we can minimize the following objective function with respect to both $\mathbf{z}_1,\ldots,\mathbf{z}_n$ and $\boldsymbol{\mu},\boldsymbol{\Omega}$,

$$\begin{aligned} \ell(\mathbf{z}_1,\ldots,\mathbf{z}_n,\boldsymbol{\mu},\boldsymbol{\Omega}) = &-\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=1}^{K} x_{i,k}z_{i,k} - M_i\log\left(\sum_{k=1}^{K} e^{z_{i,k}}+1\right)\right] \\ &-\frac{1}{2}\log[\det(\boldsymbol{\Omega})] + \frac{1}{2n}\sum_{i=1}^{n}(\mathbf{z}_i-\boldsymbol{\mu})'\boldsymbol{\Omega}(\mathbf{z}_i-\boldsymbol{\mu}) + \lambda\|\boldsymbol{\Omega}\|_1. \end{aligned} \tag{5}$$

In other words, $\mathbf{z}_1,\ldots,\mathbf{z}_n$, $\boldsymbol{\mu}$, and $\boldsymbol{\Omega}$ are all treated as unknown parameters for minimization in the objective function (5).

The objective function (5) was introduced as profiling out the parameters $(\mathbf{z}_1,\ldots,\mathbf{z}_n)$ and then minimizing the posterior distribution over $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$. However, there is a Bayesian interpretation as well. Notice that we have set up a logistic normal multinomial model as in (1)–(3). If we treat $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ as random and impose a hyperprior distribution on them as

$$\mu \sim \text{Lebesgue}(-\infty,\infty) \quad \text{and} \quad \boldsymbol{\Omega} \sim \text{Laplace}(0, 1/(n\lambda)),$$

then, (5) becomes the negative logarithm of the posterior distribution of $\mathbf{z}_1,\ldots,\mathbf{z}_n,\boldsymbol{\mu},\boldsymbol{\Omega}$ given $\mathbf{x}_1,\ldots,\mathbf{x}_n$. Therefore, minimizing (5) is equivalent to finding the maximum a posteriori (MAP) estimator of $\mathbf{z}_1,\ldots,\mathbf{z}_n,\boldsymbol{\mu},\boldsymbol{\Omega}$ given the data $\mathbf{x}_1,\ldots,\mathbf{x}_n$.

2.3 Computational Algorithm

The objective function (5) includes naturally three sets of parameters $(\mathbf{z}_1, \ldots, \mathbf{z}_n)$, $\boldsymbol{\mu}$, and $\boldsymbol{\Omega}$, which motivates us to apply a block coordinate descent algorithm. A block coordinate descent algorithm minimizes the objective function iteratively for each set of parameters given the other sets. Given the initial values $(\mathbf{z}_1^{(0)}, \ldots, \mathbf{z}_n^{(0)})$, $\boldsymbol{\mu}^{(0)}$, and $\boldsymbol{\Omega}^{(0)}$, a block coordinate algorithm repeats the following steps cyclically for iteration $t = 0, 1, 2, \ldots$ until the algorithm converges.

1. Given $\boldsymbol{\mu}^{(t)}$ and $\boldsymbol{\Omega}^{(t)}$, find $(\mathbf{z}_1^{(t+1)}, \ldots, \mathbf{z}_n^{(t+1)})$ that maximizes (5).
2. Given $(\mathbf{z}_1^{(t+1)}, \ldots, \mathbf{z}_n^{(t+1)})$ and $\boldsymbol{\Omega}^{(t)}$, find $\boldsymbol{\mu}^{(t+1)}$ that maximizes (5).
3. Given $(\mathbf{z}_1^{(t+1)}, \ldots, \mathbf{z}_n^{(t+1)})$ and $\boldsymbol{\mu}^{(t+1)}$, find $\boldsymbol{\Omega}^{(t+1)}$ that maximizes (5).

As follows, we will present the details of this algorithm in each iteration. For the initial values $(\mathbf{z}_1^{(0)}, \ldots, \mathbf{z}_n^{(0)})$, we use their maximum likelihood estimators from the multinomial distribution, i.e.,

$$z_{i,k}^{(0)} = \log\left(\frac{x_{i,k}}{x_{i,K+1}}\right), \ i = 1, \ldots, n, \ k = 1, \ldots, K.$$

If $x_{i,K+1} = 0$ for some i, we add a small constant to the denominator in the logarithm function. For the initial values $\boldsymbol{\mu}^{(0)}$, notice that we have a closed form minimizer of $\boldsymbol{\mu}$ for (5) given the values of $(\mathbf{z}_1, \ldots, \mathbf{z}_n)$, which is $\boldsymbol{\mu} = \bar{\mathbf{z}} = \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i$. Therefore, we set the initial value as $\boldsymbol{\mu}^{(0)} = \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i^{(0)}$. Finally, for the initial value $\boldsymbol{\Omega}^{(0)}$, we use the estimate of the graphical lasso algorithm taking the sample covariance matrix computed from $\mathbf{z}_1^{(0)}, \ldots, \mathbf{z}_n^{(0)}$ as input.

In step 1, given $\boldsymbol{\mu}^{(t)}$ and $\boldsymbol{\Omega}^{(t)}$, minimizing the objective function (5) with respect to $(\mathbf{z}_1, \ldots, \mathbf{z}_n)$ is equivalent to minimizing the following objective function with respect to each $\mathbf{z}_i$ separately, for $i = 1, \ldots, n$:

$$\ell_i^{(t)}(\mathbf{z}_i) = \frac{1}{2}(\mathbf{z}_i - \boldsymbol{\mu}^{(t)})'\boldsymbol{\Omega}^{(t)}(\mathbf{z}_i - \boldsymbol{\mu}^{(t)}) - \left[\sum_{k=1}^K x_{i,k} z_{i,k} - M_i \log\left(\sum_{k=1}^K e^{z_{i,k}} + 1\right)\right]. \qquad (6)$$

The above objective function is a smooth and convex function in $\mathbf{z}_i$ and its Hessian matrix is positive definite. Therefore, we apply the Newton-Raphson algorithm to find the minimizer

numerically. In addition, we implement a line search procedure in each Newton-Raphson iteration following the Armijo rule (Armijo, 1966). This procedure ensures sufficient decrease in the objective function at each iteration to prevent possible divergence of the algorithm.

Step 2 is similar to the initialization step, in which $\boldsymbol{\mu}$ has a closed-form solution and is updated as $\bar{\mathbf{z}}^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i^{(t+1)}$ from the current numerical values of $(\mathbf{z}_1^{(t+1)}, \ldots, \mathbf{z}_n^{(t+1)})$ that are computed from the Newton-Raphson algorithm in step 1.

In step 3, given $(\mathbf{z}_1^{(t+1)}, \ldots, \mathbf{z}_n^{(t+1)})$ and $\boldsymbol{\mu}^{(t+1)} = \bar{\mathbf{z}}^{(t+1)}$, the objective function for $\boldsymbol{\Omega}$ can be simplified as

$$\begin{aligned}\ell^{(t)}(\boldsymbol{\Omega}) &= -\frac{1}{2}\log[\det(\boldsymbol{\Omega})] + \frac{1}{2n}\sum_{i=1}^{n}(\mathbf{z}_i^{(t+1)} - \boldsymbol{\mu}^{(t+1)})'\boldsymbol{\Omega}(\mathbf{z}_i^{(t+1)} - \boldsymbol{\mu}^{(t+1)}) + \lambda\|\boldsymbol{\Omega}\|_1, \\ &= -\frac{1}{2}\log[\det(\boldsymbol{\Omega})] + \frac{1}{2}\mathrm{tr}(\mathbf{S}^{(t+1)}\boldsymbol{\Omega}) + \lambda\|\boldsymbol{\Omega}\|_1, \end{aligned} \tag{7}$$

where $\mathbf{S}^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{z}_i^{(t+1)} - \bar{\mathbf{z}}^{(t+1)})(\mathbf{z}_i^{(t+1)} - \bar{\mathbf{z}}^{(t+1)})'$. It is obvious that minimizing the objective function (7) becomes a graphical lasso problem (Yuan and Lin, 2007; Friedman et al., 2008). It is well known that the graphical lasso objective function is a convex function in $\boldsymbol{\Omega}$ and efficient algorithms have been developed for its optimization (Friedman et al., 2008). In this paper, we implement this step using the graphical lasso algorithm included in the `huge` (Zhao et al., 2012) package in R.

The above block coordinate descent algorithm iterates between Newton-Raphson and graphical lasso and is designed specifically to optimize the logistic normal multinomial model for compositional count data. Therefore, we name this algorithm the compositional graphical lasso algorithm, and the entire approach the compositional graphical lasso method including both the logistic normal multinomial model and the compositional graphical lasso algorithm for the analysis of compositional count data such as microbiome abundance data.

2.4 Theoretical Convergence

Unfortunately, the objective function (5) is not necessarily a convex function jointly in $(\mathbf{z}_1, \ldots, \mathbf{z}_n)$, $\boldsymbol{\mu}$, and $\boldsymbol{\Omega}$. However, we have shown that it is actually convex in each subset of its parameters (see Section 2.3). The convergence properties of such an optimization

problem has been extensively studied in the literature. For example, Tseng (2001) studied the convergence properties of a block coordinate descent method applied to minimize a nonconvex function with certain separability and regularity properties. We will establish the convergence properties of the compositional graphical lasso algorithm following Tseng (2001).

Recall that our algorithm treats the three sets of parameters $(\mathbf{z}_1, \ldots, \mathbf{z}_n)$, $\boldsymbol{\mu}$, and $\boldsymbol{\Omega}$ as three blocks and optimize for each block iteratively. In addition, as in Tseng (2001), the objective function (5) can be regarded as the sum of two parts, the first of which is an inseparable but differentiable function as

$$\ell_0(\mathbf{z}_1, \ldots, \mathbf{z}_n, \boldsymbol{\mu}, \boldsymbol{\Omega}) = \frac{1}{2n} \sum_{i=1}^{n} (\mathbf{z}_i - \boldsymbol{\mu})' \boldsymbol{\Omega} (\mathbf{z}_i - \boldsymbol{\mu}), \tag{8}$$

and the second of which is a sum of separable and differentiable functions as $\ell_1(\mathbf{z}_1, \ldots, \mathbf{z}_n) + \ell_2(\boldsymbol{\Omega})$, where

$$\ell_1(\mathbf{z}_1, \ldots, \mathbf{z}_n) = -\frac{1}{n} \sum_{i=1}^{n} \left[\sum_{k=1}^{K} x_{i,k} z_{i,k} - M_i \log \left(\sum_{k=1}^{K} e^{z_{i,k}} + 1 \right) \right], \tag{9}$$

$$\ell_2(\boldsymbol{\Omega}) = -\frac{1}{2} \log[\det(\boldsymbol{\Omega})] + \lambda \|\boldsymbol{\Omega}\|_1. \tag{10}$$

Tseng (2001) established the convergence properties of a block coordinate descent algorithm under regularity conditions on ℓ_0, ℓ_1, and ℓ_2.

To present the major convergence property of the compositional graphical lasso algorithm, let's review the definition of a cluster point in real analysis. A cluster point of a set $\mathcal{A} \subset \mathbb{R}^n$ is a real vector $\mathbf{a} \in \mathbb{R}^n$ such that for every $\delta > 0$, there exists a point $\mathbf{x}$ in $\mathcal{A} \setminus \{\mathbf{a}\}$ such that $\|\mathbf{x} - \mathbf{a}\|_2 < \delta$. Obviously, any limit point of the set $\mathcal{A}$ is a cluster point.

Furthermore, define a cluster point of the compositional graphical lasso algorithm to be a cluster point of the set $\{(\mathbf{z}_1^{(t)}, \ldots, \mathbf{z}_n^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Omega}^{(t)}) : t = 0, 1, 2, \ldots\}$, which are minimizers found at each iteration t. Then, the following theorem presents a theoretical property for every cluster point of our algorithm as follows.

Theorem 1. *Any cluster point of the compositional graphical lasso algorithm is a stationary*

point of the objective function (5).

Proof. The conclusion in Theorem 1 is directly implied by Theorem 4.1(c) in Tseng (2001), for which we need to verify a few regularity conditions as follows.

First, $\ell_0(\mathbf{z}_1, \ldots, \mathbf{z}_n, \boldsymbol{\mu}, \boldsymbol{\Omega})$ in (8) is regular at each point in its domain. This is true because $\text{dom}(\ell_0)$ is open and ℓ_0 is differentiable and all its partial derivatives exists.

Second, the level set $\{(\mathbf{z}_1, \ldots, \mathbf{z}_n, \boldsymbol{\mu}, \boldsymbol{\Omega}) : \ell(\mathbf{z}_1, \ldots, \mathbf{z}_n, \boldsymbol{\mu}, \boldsymbol{\Omega}) \leq \ell(\mathbf{z}_1^{(0)}, \ldots, \mathbf{z}_n^{(0)}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Omega}^{(0)})\}$ is compact and that ℓ in (5) is continuous on this level set. The continuity part is obvious and we just need to argue the compactness of the level set.

Let's argue this by first proving that ℓ in (5) is bounded below.

$$
\begin{aligned}
\ell = & -\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=1}^{K} x_{i,k}z_{i,k} - M_i \log\left(\sum_{k=1}^{K} e^{z_{i,k}} + 1\right)\right] \\
& - \frac{1}{2}\log[\det(\boldsymbol{\Omega})] + \frac{1}{2n}\sum_{i=1}^{n}(\mathbf{z}_i - \boldsymbol{\mu})'\boldsymbol{\Omega}(\mathbf{z}_i - \boldsymbol{\mu}) + \lambda\|\boldsymbol{\Omega}\|_1 \\
& \geq -\frac{1}{2}\log[\det(\boldsymbol{\Omega})] + \frac{1}{2n}\sum_{i=1}^{n}(\mathbf{z}_i - \boldsymbol{\mu})'\boldsymbol{\Omega}(\mathbf{z}_i - \boldsymbol{\mu}) + \lambda\|\boldsymbol{\Omega}\|_1 \\
& \geq -\frac{1}{2}\log[\det\left(\hat{\boldsymbol{\Omega}}\right)] + \frac{1}{2n}\sum_{i=1}^{n}(\mathbf{z}_i - \boldsymbol{\mu})'\hat{\boldsymbol{\Omega}}(\mathbf{z}_i - \boldsymbol{\mu}) + \lambda\|\hat{\boldsymbol{\Omega}}\|_1 \\
& \geq -\frac{1}{2}\log[\det\left(\hat{\boldsymbol{\Omega}}\right)] \\
& \geq -\frac{1}{2}K\log\frac{K}{\lambda}, \\
& \text{where } \hat{\boldsymbol{\Omega}} = \arg\min_{\boldsymbol{\Omega}} -\frac{1}{2}\log[\det\left(\hat{\boldsymbol{\Omega}}\right)] + \frac{1}{2n}\sum_{i=1}^{n}(\mathbf{z}_i - \boldsymbol{\mu})'\hat{\boldsymbol{\Omega}}(\mathbf{z}_i - \boldsymbol{\mu}) + \lambda\|\hat{\boldsymbol{\Omega}}\|_1.
\end{aligned}
$$

Here, $\hat{\boldsymbol{\Omega}}$ is unique, and has positive eigenvalues bounded above by $\frac{K}{\lambda}$ (Banerjee et al., 2008).

Therefore, the level set can be written as $\{(\mathbf{z}_1, \ldots, \mathbf{z}_n, \boldsymbol{\mu}, \boldsymbol{\Omega}) : c_1 \leq \ell(\mathbf{z}_1, \ldots, \mathbf{z}_n, \boldsymbol{\mu}, \boldsymbol{\Omega}) \leq c_2\}$ for some constant c_1 and c_2, thus is compact as it is a preimage of a compact set under a continuous function.

Third, $\ell(\mathbf{z}_1, \ldots, \mathbf{z}_n, \boldsymbol{\mu}, \boldsymbol{\Omega})$ has at most one minimum in its second block of parameters, i.e., $\boldsymbol{\mu}$. This is true given that the Hessian matrix for $\boldsymbol{\mu}$ is $\boldsymbol{\Omega}$ which is positive definite.

The conclusion of Theorem 1 is proved as we have verified all regularity conditions in Theorem 4.1(c) in Tseng (2001). □

It is also noteworthy that the values of the objective function at each iteration, i.e., $\{\ell(\mathbf{z}_1^{(t)}, \ldots, \mathbf{z}_n^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Omega}^{(t)}) : t = 0, 1, 2, \ldots\}$ will always converge. This is because that the objective function is bounded below and our algorithm results in non-increasing objective function values between two iterations. Therefore, the values of objective function will always converge to a limit point. However, this does not guarantee the convergence of the minimizers, i.e., $\{(\mathbf{z}_1^{(t)}, \ldots, \mathbf{z}_n^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Omega}^{(t)}) : t = 0, 1, 2, \ldots\}$. Instead, Theorem 1 provides some theoretical guarantees about the convergence of the minimizers, which states that any cluster point of the algorithm is a stationary point.

In practice, we have always observed the numerical convergence of the minimizers after a certain number of iterations. Therefore, Theorem 1 guarantees that the solution from the algorithm is at least a stationary point. To achieve global optimization, one can run the algorithm multiple times starting with different initial values and choose the one solution that yields the smallest objective function among the multiple ones.

2.5 Tuning Parameter Selection

There is a large body of literature on the selection a tuning parameter in the variable selection framework. Parameter selection approaches include criterion-based methods such as Akaike information criterion (AIC) (Akaike, 1974) and Bayesian information criteria (BIC) (Schwarz et al., 1978) that balance the model complexity and the goodness of fit, prediction-based methods such as cross validation (Larson, 1931; Mosteller and Wallace, 1963; Mosteller and Tukey, 1968; Stone, 1974; Geisser, 1975) and generalized cross validation (Golub et al., 1979) that aim to minimize the expected prediction error of the selected model on independent datasets, and stability-based methods such as stability selection (Meinshausen and Bühlmann, 2010) and Stability Approach to Regularization Selection (StARS) (Liu et al., 2010) that select a model with high stability under subsampling or bootstapping the original data.

In this work, we applied StARS to select the tuning parameter λ in our objective function

(5). In StARS, we draw N subsamples without replacement from the original dataset with n observations, each of size b. For each tuning parameter λ, we obtain an estimate of $\mathbf{\Omega}$, i.e., a network for each subsample. Then, we measure the total instability of these resultant networks across the N subsamples. The total instability of these networks is defined by averaging the instabilities of each edge across the N subsamples over all possible edges, where the instability of each edge is estimated as the twice the sample variance of the Bernoulli indicator of whether this edge is selected or not in these N subsamples.

The tuning parameter is then selected to be the maximum value of λs with which the instability of the resultant networks is less than a threshold $\beta > 0$. In principle, StARS selects a tuning parameter so that the resultant network is the sparsest among all networks with a total stability less than a threshold β.

3 Simulation Study

3.1 Settings

To evaluate the performance of compositional graphical lasso, we conduct a simulation study and compare it with other network estimation methods such as neighborhood selection (Meinshausen et al., 2006) and graphical lasso (Friedman et al., 2008).

Given our goal is to estimate the true network, i.e., $\mathbf{\Omega}$ in (3), we consider three types of precision matrix $\mathbf{\Omega} = (\omega_{kl})_{1 \leq k,l \leq K}$, which are different in the pattern of edge distributions as well as the degree of connectedness.

1. Chain: $\omega_{kk} = 1.5$, $\omega_{kl} = 0.5$ if $|k - l| = 1$, and $\omega_{kl} = 0$ if $|k - l| > 1$. A node is designed to be connected to its adjacent nodes, and the connectedness of nodes is balanced.

2. Random: $\omega_{kl} = 1$ with probability $3/K$ for $k \neq l$. A node is connected to all other nodes randomly with a fixed probability. Similar to the chain structure, the connectedness of nodes is balanced.

3. Hub: All nodes are randomly split into $\lceil K/20 \rceil$ disjoint groups, and a hub node k is selected from each group. For any other node l in the same group, $\omega_{kl} = 1$. All the

remaining entries of $\mathbf{\Omega}$ are zero. Here, nodes are partitioned into the same group at random, but is then designated to be connected to the hub node at certain. The degree of connectedness among nodes is extremely unbalanced in this case: the hub nodes are connected to all the other nodes in its group (around 20 nodes) and all the other nodes are only connected to the hub node in its group, i.e., just one node.

In addition to the true network, we also consider two other factors that are expected to influence the result. The first factor is the sequencing depth, M_i, in the multinomial distribution (1). We simulate M_i from two uniform distributions, Unif$(20K, 40K)$ and Unif$(100K, 200K)$, and call the two settings low and high sequencing depth, respectively. The second factor is the variation included in the logistic normal distribution (3). Although we consider three types of precision matrices, we consider an additional factor by multiplying a positive constant c to $\mathbf{\Omega}$ so that the true precision matrix is $c\mathbf{\Omega}$. We choose $c = 1$ and $c = 1/5$ separately and call the two setting low and high compositional variation, respectively.

The data are simulated following the logistic normal multinomial model in (1)–(3). We first simulate $\mathbf{z}_i \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$ independently for $i = 1, \ldots, n$; then, we perform the inverse log-ratio transformation (also know as the softmax transformation, the inverse transformation of (2)) to obtain the multinomial probabilities $\mathbf{p}_i$ for $i = 1, \ldots, n$; last, we simulate multinomial counts $\mathbf{x}_i$ from a multinomial distribution with sequencing depth M_i and probabilities $\mathbf{p}_i$. Throughout this simulation study, we fix $n = 100$ and $K = 200$.

The simulation results are based on 100 replicates of simulated data. On each replicate, we apply compositional graphical lasso, neighborhood selection, and graphical lasso separately to obtain a sparse estimator of $\mathbf{\Omega}$. For neighborhood selection and graphical lasso, we first obtain an estimate of $\mathbf{z}_1, \ldots, \mathbf{z}_n$ from the multinomial distribution as

$$\tilde{z}_{i,k} = \log\left(\frac{x_{i,k}}{x_{i,K+1}}\right), \ i = 1, \ldots, n, \ k = 1, \ldots, K, \tag{11}$$

and then apply neighborhood selection and graphical lasso directly on the estimates $\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_n$ by treating them as surrogates for their true counterparts, i.e., $\mathbf{z}_1, \ldots, \mathbf{z}_n$.

To compare the performance of the three methods in terms of network recovery, all three methods are applied with a sequence of 70 tuning parameter values, and their true positive

rates (TPR) and false positive rates (FPR) in terms of edge selection are recorded for each value of λ. An ROC curve is plotted from the average TPR and the average FPR over the 100 replicates at each position of the tuning parameter sequence.

In addition, we apply StARS to select an optimal tuning parameter λ. Following the recommendation in Liu et al. (2010), we set the threshold for the total instability to be $\beta = 0.05$, the size of each subsample $b = 7\sqrt{n}$, and the number of subsamples $N = 50$. Once the optimal tuning parameter is determined by StARS, we fit the whole dataset with the selected tuning parameter and evaluate the resultant network using three criteria: precision, recall, and F1 score, which are defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}},$$

where TP, FP, and FN are numbers of true positives, false positives, and false negatives, respectively.

3.2 Results

Figure 1 presents the ROC curves for compositional graphical lasso (Comp-gLASSO), neighborhood selection (MB), and graphical lasso (gLASSO), from which we can see that compositional graphical lasso dominates its competitors in terms of edge selection in all settings. In particular, the advantage of the compositional graphical lasso over neighborhood selection, and graphical lasso is the most obvious when the compositional variation is high and the sequencing depth is low, no matter which type of network structure is considered. On the contrary, the three methods perform very similarly for all types of network structures when the compositional variation is low and the sequencing depth is high. The difference between compositional graphical lasso and the rest is intermediate for the other two settings when both compositional variation and sequencing depth are high or low. Comparing graphical lasso and neighborhood selection, they tend to perform more similarly although graphical lasso seems to outperform neighborhood selection in some settings with a small margin.

The above observations agree with our expectation about how the two factors, compositional variation and sequencing depth, affect the simulation results. Recall that neigh-

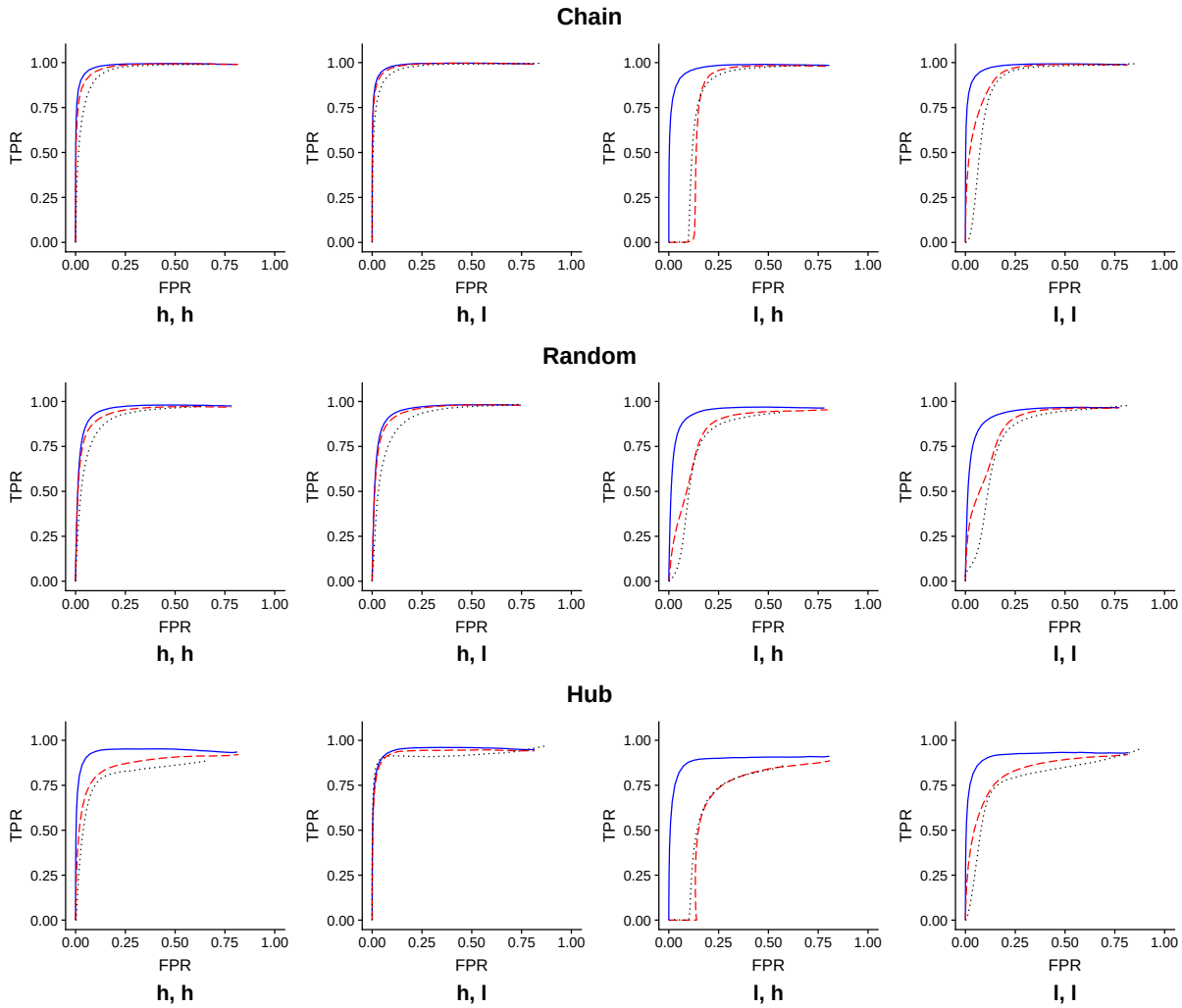


Figure 1: ROC curves for compositional graphical lasso (Comp-gLASSO), graphical lasso (gLASSO) and neighborhood selection (MB). Solid blue: Comp-gLASSO; dashed red: gLASSO; dotted black: MB. $\mathbf{h}, \mathbf{h}$: high sequencing depth and high compositional variation; $\mathbf{h}, \mathbf{l}$: high sequencing depth and low compositional variation; $\mathbf{l}, \mathbf{h}$: low sequencing depth and high compositional variation; $\mathbf{l}, \mathbf{l}$: low sequencing depth and low compositional variation.

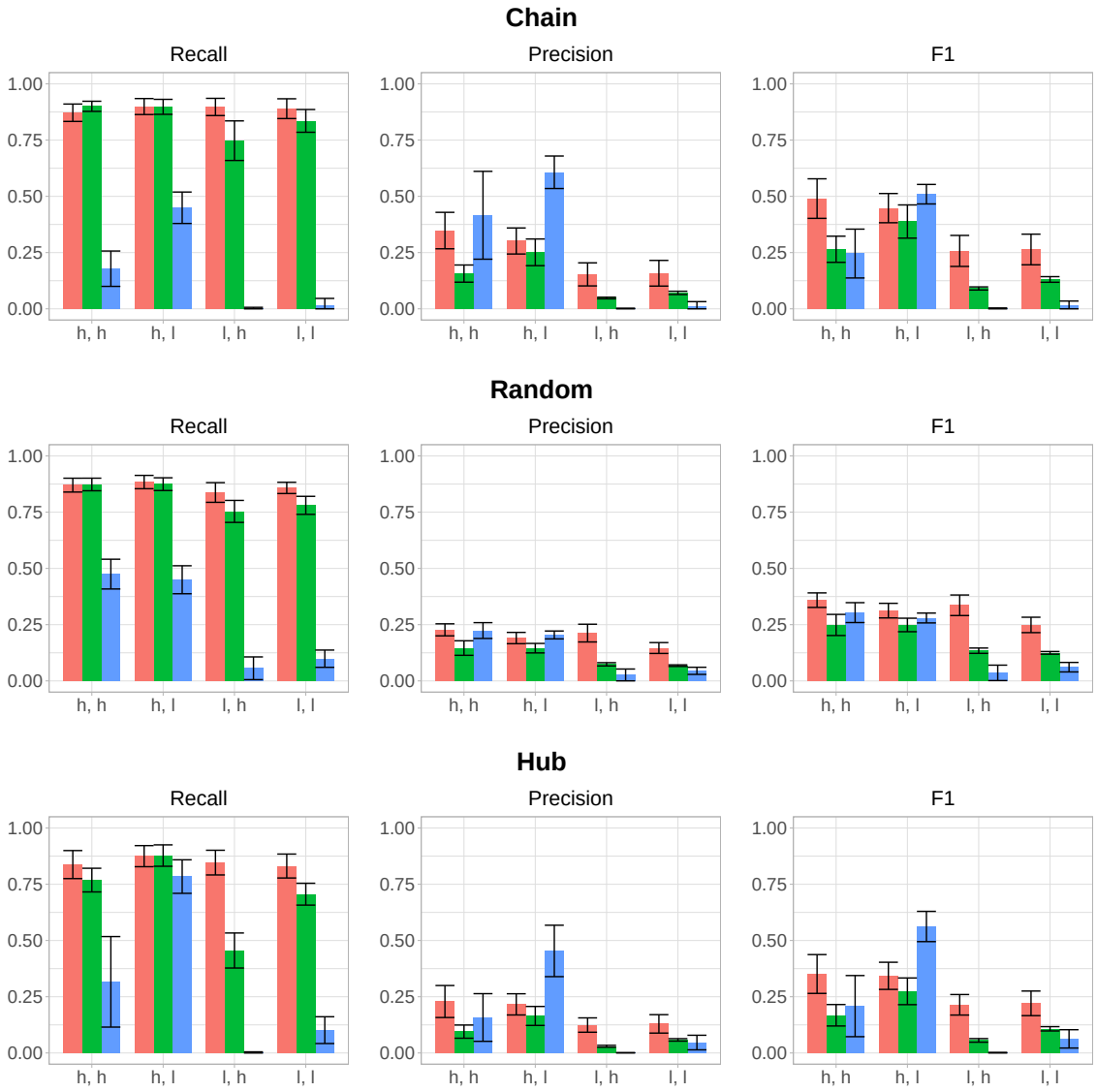


Figure 2: Recall, precision and F1 score for the network selected by StARS for compositional graphical lasso (Comp-gLASSO), graphical lasso (gLASSO) and neighborhood selection (MB). Red (left): Comp-gLASSO; green (middle): gLASSO; blue (right): MB. $\mathbf{h}, \mathbf{h}$: high sequencing depth and high compositional variation; $\mathbf{h}, \mathbf{l}$: high sequencing depth and low compositional variation; $\mathbf{l}, \mathbf{h}$: low sequencing depth and high compositional variation; $\mathbf{l}, \mathbf{l}$: low sequencing depth and low compositional variation.

borhood selection and graphical lasso replace the true values of $\mathbf{z}_1, \ldots, \mathbf{z}_n$ by their estimates/surrogates $\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_n$ as in (11) without taking into account the estimation accuracy or uncertainty of these surrogates. On the one hand, a higher sequencing depth leads to more accurate surrogates $\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_n$; therefore, it is not surprising to see that all three methods perform more similarly when the sequencing depth is high. On the other hand, a higher compositional variation results in a higher variation in $\mathbf{z}_i$'s and further in $\mathbf{p}_i$'s that are multinomial probabilities. Since neighborhood selection and graphical lasso ignore the multinomial component in the model, it is also not surprising to see that their performances are deteriorated by a high compositional variation.

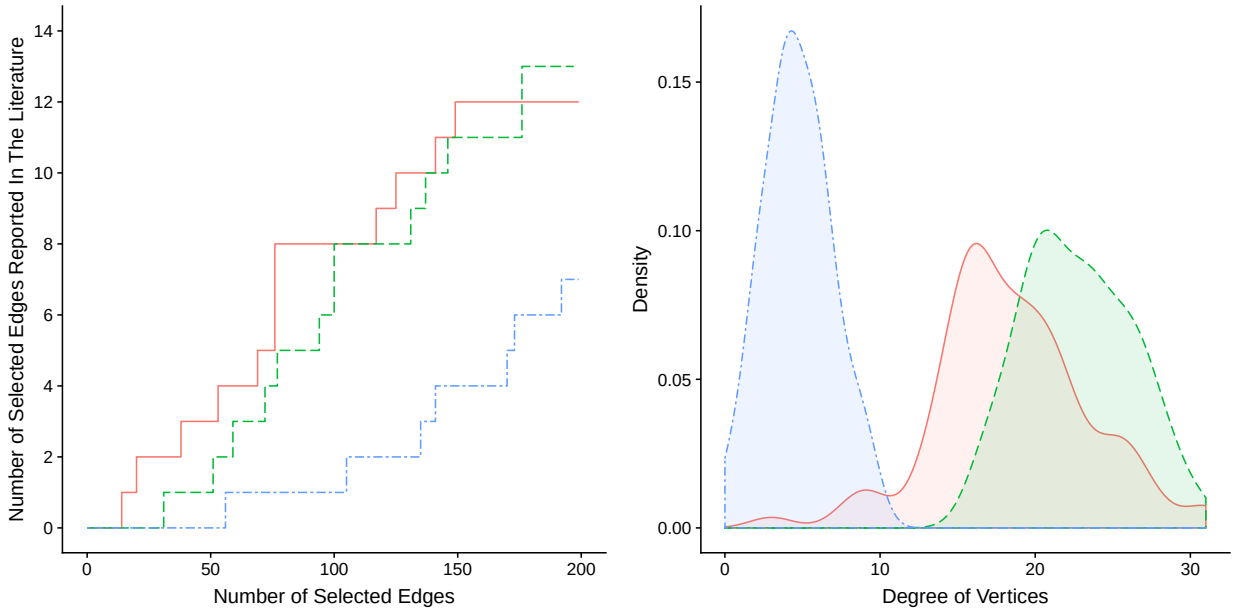
Figure 2 presents recall, precision, and F1 score from 50 replicates of the estimated network resulted from the tuning parameter selected by StARS. The first observation would be that the precisions of both compositional graphical lasso and graphical lasso are much worse than their recalls, due to the high-level sparsity in the precision matrix, whereas the precisions and recalls are more comparable for neighborhood selection. Interestingly, StARS results in a much more sparse network for neighborhood selection than the other methods under the same stability threshold, suggesting that fewer edges selected by neighborhood selection are stable enough (passing the stability threshold). When it comes to method comparison, compositional graphical lasso has much higher recall than neighborhood selection in most settings, but have comparable or lower recall in most of the settings with high sequencing depth. The network from compositional graphical lasso has higher F1 score than the ones from neighborhood selection in most settings, except when sequencing depth is high and compositional variation is low for chain and hub networks. In addition, the network from compositional graphical lasso has higher precision, recall, and F1 score than the ones from graphical lasso in almost all settings, with the only exception of a lower recall when both compositional variation and sequencing depth are high for chain and random networks. Similar to the observations from the ROC curves, the advantage of compositional graphical lasso is more obvious with a low sequencing depth or a high compositional variation.

4 Real Data

To better understand the ocean, the largest ecosystem on the earth, the Tara Oceans consortium sampled both plankton and environmental data in 210 sites from the world oceans, using the 110-foot research schooner Tara during the 2009-2013 Tara Oceans expedition (2009-2013). The data collected was later analyzed using sequencing and imaging techniques. As part of the TARA Oceans project, Lima-Mendez et al. (2015) analyzed the interactions between oceanic microbes, and provided a list of 91 gold-standard genus-level marine planktonic interactions that are described in the literature. Though this list only comprises interactions between a small fraction of the total marine eukaryotic diversity and is therefore far from complete, it could serve as partial ground truth to evaluate the interactions identified by different methods. We downloaded the taxonomic data and the literature interactions from the TARA Ocean Project data repository (`https://doi.pangaea.de/10.1594/PANGAEA.843018`).

As the partial ground truth is a list of genus-level interactions, we choose to analyze the genus-level abundance data, which are aggregated from the original OTU abundance data. To reduce the computational complexity, we only include the 81 genera that are involved in the list of gold-standard interactions in our analysis. In addition, we discard the samples with too few sequence reads (less than 100), resulting in 324 samples left in our analysis. Therefore, the final genus abundance data in our analysis has 324 samples and 81 genera.

Similar to the simulation study, we apply compositional graphical lasso, graphical lasso, and neighborhood selection to estimate the interaction network among the 81 genera. To this end, we first pick genus Acrosphaera, which was not involved in the gold-standard list that has the largest average relative abundance among those genera and use this genus as the reference taxon for all three methods. Then, we apply each method with a sequence of 70 decreasing tuning parameter values, resulting a sequence of interaction networks starting from an empty network, to compare the performance of compositional graphical lasso, graphical lasso and neighborhood selection. Finally, we apply StARS to find the optimal tuning parameter, in which the parameters β, b, and N are set the same as in the simulation study. We reported the final interaction networks estimated by the three algorithms fitting on the whole dataset with the optimal tuning parameters selected by StARS.



(a) (b)

Figure 3: (a): Number of identified literature interactions versus number of edges of the estimated network from the TARA dataset. (b): The degree distribution of vertices from the networks selected by StARS. Solid red: compositional graphical lasso; dashed green: graphical lasso; dashed dotted blue: neighbor hood selection.

First, to compare the three methods in terms of their ability to reconstruct the literature interactions, we apply each of them on the decreasing sequence of penalty parameters and report the number of literature interactions selected by each method among the top ranked edges. In detail, we start with a large tuning parameter that results in an empty network, then decrease the tuning parameter so that the network becomes denser, and stop until the network has about 200 edges (out of a total of 3240 possible edges). At each tuning parameter, we plot the number of literature interactions included in the network versus the total number of edges of the network, resulting in a step-function shaped curve for each method as in Figure 3a.

From Figure 3a, we can observe that compositional graphical lasso identifies slightly more literature interactions than graphical lasso until the total number of edges arrives 175 and graphical lasso identifies one more literature interaction afterwards. Neighborhood selection selects much fewer literature interactions than either compositional graphical lasso or graphical lasso. These observations imply that compositional graphical lasso slightly

outperforms graphical lasso in reconstructing the literature interactions, while its advantage over neighborhood selection is much more obvious.

Second, for the final interaction networks with the optimal tuning parameters selected by StARS, we find that compositional graphical lasso, graphical lasso, and neighborhood selection identify 749, 921 and 190 edges, respectively, with the same instability threshold used in StARS. This agrees with our observation in the simulation study that the network from neighborhood selection is much sparser than those from compositional graphical lasso and graphical lasso. The degree distributions from the networks estimated by the three methods are shown in Figure 3b. It looks like the degree distribution of the network from neighborhood selection is highly right-skewed, and the one from graphical lasso is quite left-skewed, while the one from compositional graphical lasso is relatively symmetric, though still slightly left-skewed. The center of the three degree distributions are ranked as neighborhood selection, compositional graphical lasso and graphical lasso in ascending order, which also reflects that the densities of the selected networks are in this order.

It is observed that there are a few hub genera that have an excessive number of interactions with other genera reported in the literature, such as Amoebophrya, Blastodinium and Parvilucifera. Though this literature-reported graph structure is rather incomplete, it is still of interest to evaluate how well the three methods pick up those hub genera. Since the density of networks from the three methods are rather different, it is hard to compare the the degrees of those seven hubs from the three networks directly, but it is reasonable to compare the rank of those degrees within each degree distributions. The method which has lower ranks (degree of genera ranked in descending order) for those hubs in their degree distribution are believed to pick up the hub genera better. A list of 7 hubs (which has degree ≥ 5) along with their degrees from the incomplete graph constructed from literature is shown in 1, followed by the corresponding ranks of those genera in the degree distributions from each of the three methods and their degrees in the three estimated networks in the parentheses. We can see that compositional graphical lasso has lower rank than graphical lasso for all 7 genera, while neighborhood selection has lower ranks than compositional graphical lasso for 3 genera, and the opposite for the other 4 genera. Overall, compositional graphical lasso performs the best in picking-up the literature reported hub genera among the three.

	Comp-gLASSO	gLASSO	MB
Amoebophrya(21)	31(19)	47(21)	2(9)
Blastodinium(12)	13(23)	26(24)	30(5)
Parvilucifera(12)	46(17)	67(19)	58(3)
Syndinium(7)	14(23)	27(24)	19(6)
Vampyrophrya(7)	34(19)	60(20)	6(8)
Phaeocystis(6)	1(31)	3(29)	17(6)
Pirsonia(5)	64(15)	68(19)	33(5)

Table 1: For the "hub" genera in literature, their rank in the degree distributions (in descending order) from compositional graphical lasso (Comp-gLASSO), graphical lasso (gLASSO) and neighborhood selection (MB). The numbers in the parentheses are the degrees of the genera. In particular, the numbers in the parentheses from the first columns are the degrees of genera from the literature list.

To further compare these networks with the literature interactions, we visualize these networks in accompany with the network of the literature interactions. For better visualization, we only keep the top 100 edges that are ranked by the following two criteria: (a) selection probability of an edge, the proportion of times that an edge is selected from the N subsamples in StARS and (b) edge weight, the absolute value of the partial correlation that is defined as $|\hat{\omega}_{ij}|/\sqrt{\hat{\omega}_{ii}\hat{\omega}_{jj}}$ where $\hat{\omega}_{ij}$ is the (i, j) entry of the estimated inverse covariance matrix $\hat{\Omega}$. Specifically, the edges are first ranked by selection probability, and the edges with the same selection probabilities are then ranked by weight. For the networks from all three methods, darker blue means higher magnitude in partial correlation.

We can see from Figure 4 that, though still different, the networks estimated by the three algorithms have obvious similarity in the predicted edges and the strength of edges, like the genus pair "Centropages Thalassicolla" and "Acanthometra Hexaconus" are in the top 100 list from all three methods in dark blue. On the other hand, there's very few overlaps between those top 100 edges and the known interactions from literature. Since our current knowledge of the genus-level interactions are still limited, the edges that have not been reported from literature but enjoys higher stability and larger weight might suggest promising new directions for future biological investigation. There is actually 39 common edges from the the top 100 lists from the three estimated networks, and we further ranked them by: (a) the summation of selection probabilities from the three networks and (b) the summation of weights estimated from the three methods, and provided the list of the top 15

genus pairs in Appendix for the interested readers.

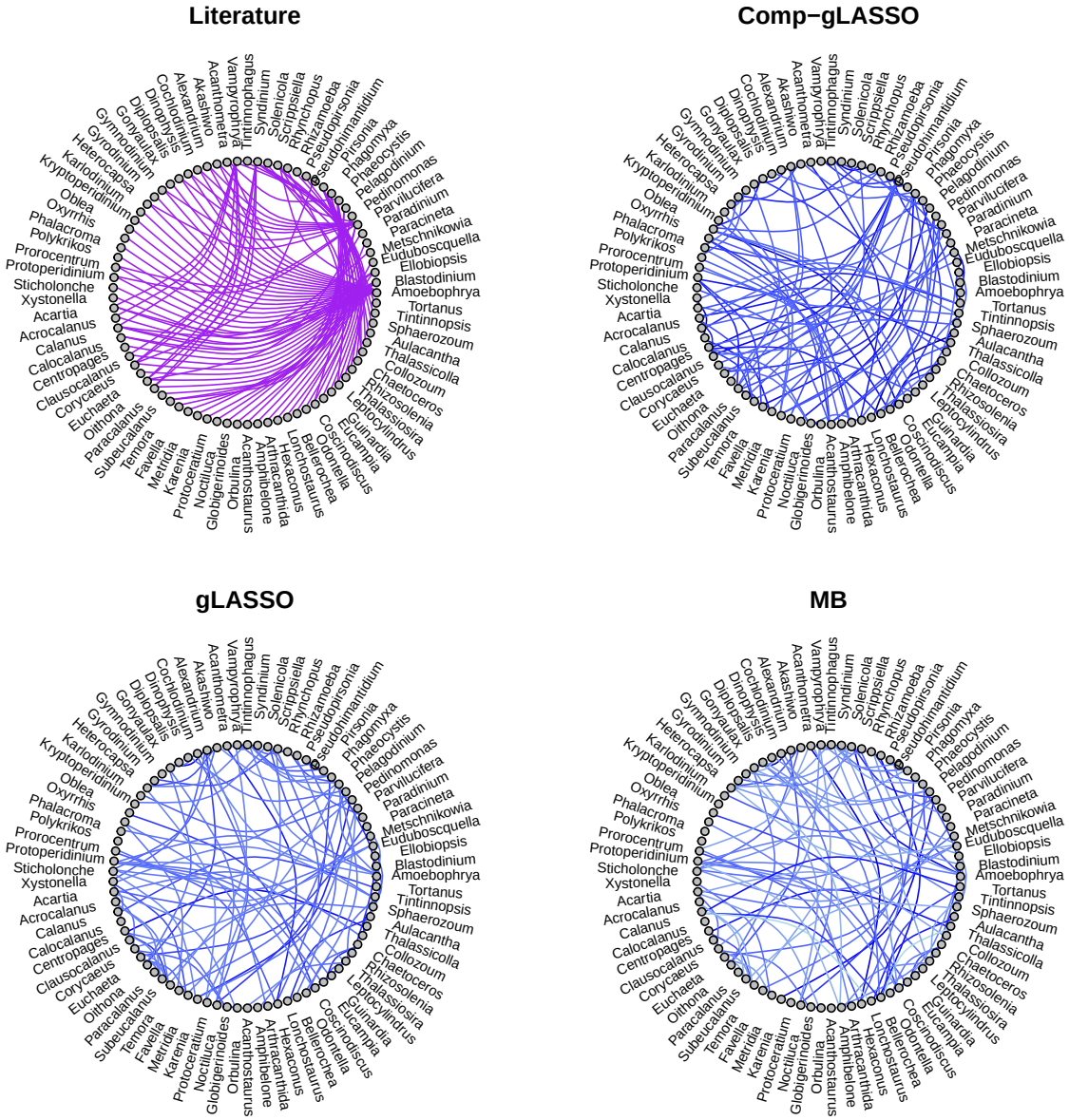


Figure 4: Inferred networks from each methods with edges filtered by selection probability and ranked by weight, compared with 91 interactions reported by literature. In the plots from all 3 methods, darker blue means stronger (large in magnitude) weight.

5 Discussion

In this work, we proposed compositional graphical lasso as a tool to estimate sparse interaction network for compositional count data based on a hierarchical model. In addition,

we have established the theoretical convergence of the algorithm. However, the theoretical property of the estimator from the algorithm still needs to be investigated. We also demonstrated the advantage of our method over other methods in multiple simulation scenarios, and applied our method to a dataset from TARA Oceans Project.

Also, though enjoying the benefits of having a full-rank precision matrix in our model, compositional graphical lasso does require one to choose a reference taxon in the first step. As a general recommendation, we suggest the readers to choose the taxon which has the highest average relative abundance among the ones they're not investigating (in reality, the number of taxa available in datasets is often far more beyond the scope that the researchers are interested about at a particular time). Since the counts of the reference taxon serve as the common denominator in the log-ratio transformation, one could be less susceptible to the problem caused by the sparsity in the denominator this way (though the undefined ratio problem could be safeguard against by adding an offset, e.g. Laplace smoothing to the data).

Readers may also wonder how much the different choices of reference taxon may change the nature of the estimated network, and if some robustness could be guaranteed across the choices of the reference. This is actually an important aspect of our current work, and a series of invariance properties regardless of the choices of references have been established. We believe that a report with theoretical investigations and analysis of synthetic and real data will come up soon.

References

Aitchison, J. (1986), *The statistical analysis of compositional data*, Monographs on statistics and applied probability, Chapman and Hall.

Akaike, H. (1974), "A new look at the statistical model identification," *IEEE transactions on automatic control*, 19, 716–723.

Armijo, L. (1966), "Minimization of functions having Lipschitz continuous first partial derivatives," *Pacific Journal of mathematics*, 16, 1–3.

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

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

Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. (2008), "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data," *Journal of Machine learning research*, 9, 485–516.

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

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.

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.

Geisser, S. (1975), "The predictive sample reuse method with applications," *Journal of the American statistical Association*, 70, 320–328.

Golub, G. H., Heath, M., and Wahba, G. (1979), "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, 21, 215–223.

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.

Larson, S. C. (1931), "The shrinkage of the coefficient of multiple correlation." *Journal of Educational Psychology*, 22, 45.

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., Chaffron, S., Ignacio-Espinosa, J. C., Roux, S., Vincent, F., et al. (2015), "Determinants of community structure in the global plankton interactome," *Science*, 348, 1262073.

Liu, H., Roeder, K., and Wasserman, L. (2010), "Stability approach to regularization selection (stars) for high dimensional graphical models," in *Advances in neural information processing systems*, pp. 1432–1440.

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

Meinshausen, N. and Bühlmann, P. (2010), "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72, 417–473.

Meinshausen, N., Bühlmann, P., et al. (2006), "High-dimensional graphs and variable selection with the lasso," *The annals of statistics*, 34, 1436–1462.

Mosteller, F. and Tukey, J. W. (1968), "Data analysis, including statistics," *Handbook of social psychology*, 2, 80–203.

Mosteller, F. and Wallace, D. L. (1963), "Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers," *Journal of the American Statistical Association*, 58, 275–309.

Schwarz, G. et al. (1978), "Estimating the dimension of a model," *The annals of statistics*, 6, 461–464.

Stone, M. (1974), "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society: Series B (Methodological)*, 36, 111–133.

Tseng, P. (2001), "Convergence of a block coordinate descent method for nondifferentiable minimization," *Journal of optimization theory and applications*, 109, 475–494.

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.

Yang, Y., Chen, N., and Chen, T. (2016), "mLDM: a new hierarchical Bayesian statistical model for sparse microbioal association discovery," *bioRxiv*, 042630.

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.

Appendix

	Genus Pair
1	Gonyaulax Alexandrium
2	Hexaconus Acanthometra
3	Thalassicolla Centropages
4	Sphaerozoum Collozoum
5	Pedinomonas Karenia
6	Phagomyxa Noctiluca
7	Orbulina Globigerinoides
8	Temora Centropages
9	Paracineta Euchaeta
10	Eucampia Coscinodiscus
11	Scrippsiella Gyrodinium
12	Vampyrophrya Syndinium
13	Tintinnophagus Rhynchopus
14	Prorocentrum Hexaconus
15	Prorocentrum Euduboscquella

Table 2: Top 15 genus pair that were predicted by compositional graphical lasso, graphical lasso and neighborhood selection, and wasn't in the literature list. All genus pairs have selection probability 1, and are then ranked by summation of absolute value in partial correlation.