AN ABSTRACT OF THE DISSERTATION OF

Chuan Tian for the degree of Doctor of Philosophy in Statistics presented on
June 12, 2020.

Title: Microbial Network Recovery by Compositional Graphical Lasso Under Additive Log-Ratio Transformation

Abstract approved: ______________________________

Yuan Jiang

The interactions between microbial taxa have been under great research interest in the science community given the microbiome data deluge. Several methods have been proposed to model and estimate the conditional dependency between microbial taxa for their interactions, in order to eliminate spurious correlation detections. However, these methods either do not account for the compositional count nature of microbiome data (such as graphical lasso), or are built upon the central log-ratio transformation (such as SPIEC-EASI) that results in a degenerate covariance matrix and thus an undefined precision matrix to present the underlying network. In addition, most existing methods ignore the potential consequence of the heterogeneity nature of microbiome data that the sum of the counts within each sample, termed "sequencing depth", can vary drastically across samples. To address these issues, we propose a novel method called "compositional graphical lasso" to identify the microbial interactions by adopting a logistic normal multinomial model which explicitly incorporates the sequencing depths. Different from most existing methods, compositional graphical lasso is based on the additive log-ratio transformation, which first selects a reference taxon and then computes the log ratios of the abundances of all the other taxa with respect to that of the reference. One natural concern about the additive log-ratio transformation would be whether the estimated network is invariant with respect to the choice of the reference. To further address this concern, we establish the reference-invariance property of a subnetwork of interest based on the additive log-ratio transformed data, and propose a reference-invariant version of the compositional graphical lasso by modifying the penalty term in its objective function to penalize only the invariant subnetwork. We illustrate the advantages of the proposed methods over the existing ones under a variety of simulation scenarios and also demonstrate their efficacy by applying them to an oceanic microbiome data set.

Microbial Network Recovery by Compositional Graphical Lasso Under Additive Log-Ratio Transformation

by

Chuan Tian

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented June 12, 2020
Commencement June 2021

Doctor of Philosophy dissertation of Chuan Tian presented on June 12, 2020.

APPROVED:

Major Professor, representing Statistics

Chair of the Department of Statistics

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Chuan Tian, Author

ACKNOWLEDGEMENTS

First and foremost, my deepest gratitude goes to my PhD advisor and the P.I. (Principle Investigator) of my research group, Dr. Yuan Jiang, for the tremendous time and efforts he invested in my growth. Being one of the most intelligent people that I know, Yuan respects me as his fellow researcher from day 1, encouraging me to generate my own research ideas, and is very open-minded about different opinions. Yuan is so generous with his time, that I could often get a short meeting with him by stopping by his open door (before the pandemic takes place, of course). Moreover, Yuan cares about his students more than professionally as colleagues, but also as people in his life. I feel comfortable talking about not only research but also school and life in general in front of Yuan, and receive helpful suggestions from him all the time. I feel blessed to have been working with him throughout my PhD program, and to have him as a friend that I look up to.

Secondly, my great appreciation of the two co-P.I.'s in my research group, Dr. Duo Jiang and Dr. Tom Sharpton, for the inspiring discussions in our group meetings, and the experiences to work together on research papers. I learned so much from them on both statistics and biology, and even more importantly, how to communicate and collaborate with people from a different academic background.

Also, I want to acknowledge Dr. Rebecca Hutchinson, Dr. Charlotte Wickham and Dr. Harold Bae for their time and contributions as members of my committee. I would like to thank them for being incredibly responsive and flexible, especially under these unusual circumstances of a pandemic. Charlotte is probably the professor I took most courseworks (Regression, Time Series, and Data Visualization) from at Oregon State University, and

Data Visualization is one of the coolest classes I have ever had in my life! I cannot express enough how much I'm indebted to her for teaching me to use *tidyverse* since my second year, which turns out to be so helpful in both presenting research works and in job interviews.

Department of Statistics at Oregon State University has been my second home in the past six years, nurturing me with the great education and the freedom to pursue my interests in addition to the required curriculum. I would like to thank both Prof. Virginia Lesser and Dr. Lisa Ganio for their leadership and genuine interest in their students' safety and success, during their time as the chair of our department. I'm also grateful to Prof. Lan Xue for her dedicated service as associate chair, which involves accommodating the complex preference requests of TA assignments from both faculty and students. Thank Maggie Neel and Mary Gardner for their extraordinary services when they were in the Statistics Main Office. Their professionalism, kindness and creativity made our great department an even warmer place. Especially, I would like to thank Dr. Lisa Madsen, who arranged my recruitment as a PhD student, and guided me through my first a couple of years in the program. I cannot remember how many times our conversations in her basement office had helped me carry through.

I was so fortunate for the friends I have made, from both Statistics and Computer Science departments. Thank you for sharing your knowledge, your enthusiasm and your wisdom with me. Even years later, I shall reflect in fond of the coffee time downtown with Chris Comiskey, the walks to Bald Hill with Chris Wolf, the hours-long conversations at KEC atrium with Anurag Koul, and so many other good times. Most of those endearing memories are shared with my best friend Zheng Liu, who stands by me in good and hard times, and is confident in me when I struggle.

Last but not least, this dissertation wouldn’t exist without my parents’ unlimited support and unreserved love. Looking forward to the day that we all get through this pandemic and see each other, once again in person.

TABLE OF CONTENTS

Page

1 Introduction 1

1.1 Challenges in Recovering Microbial Interactions 1

1.1.1 Compositionality 2

1.1.2 Finite and Heterogeneous Sequencing Depth 3

1.1.3 High-dimensionality 4

1.2 Previous Work on Microbial Interaction Networks 5

1.2.1 Marginal Correlation Networks 5

1.2.2 Conditional Dependence Networks 7

1.3 Our Contribution 9

2 Microbial Network Recovery by Compositional Graphical Lasso 11

2.1 Abstract 11

2.2 Introduction 12

2.3 Compositional Graphical Lasso 14

2.3.1 Logistic Normal Multinomial Model 14

2.3.2 Objective Function 16

2.3.3 Computational Algorithm 18

2.3.4 Theoretical Convergence 21

2.3.5 Tuning Parameter Selection 24

2.4 Simulation Study 26

2.4.1 Settings 26

2.4.2 Results 29

2.5 Real Data 33

2.6 Discussion 40

2.7 Appendix 41

3 Reference-Invariance Property of Inverse Covariance Matrix Estimation Under Additive Log-Ratio Transformation and Its Application to Microbial Network Recovery 42

3.1 Abstract 42

3.2 Introduction 43

3.3 Methodology 46

TABLE OF CONTENTS (Continued)

Page

3.3.1 Reference-Invariance Property 46
3.3.2 Logistic Normal Multinomial Model 48
3.3.3 Reference-Invariant Objective Function 50

3.4 Simulation Study 53
3.4.1 Settings 53
3.4.2 Results 56
3.4.2.1 Normalized Manhattan Similarity 56
3.4.2.2 Jaccard Index and Normalized Hamming Similarity . . . 59
3.4.2.3 ROC Curves 62

3.5 Real Data 64

3.6 Discussion 68

4 Summary and Discussion 70

4.1 Summary 70

4.2 Discussion 71
4.2.1 Computation Time 71
4.2.2 The advantage of the ALR Transformation 72
4.2.3 Alternative Approach 72

LIST OF FIGURES

Figure Page

2.1 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. **h**, **h**: high sequencing depth and high compositional variation; **h**, **l**: high sequencing depth and low compositional variation; **l**, **h**: low sequencing depth and high compositional variation; **l**, **l**: low sequencing depth and low compositional variation. 30

2.2 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. **h**, **h**: high sequencing depth and high compositional variation; **h**, **l**: high sequencing depth and low compositional variation; **l**, **h**: low sequencing depth and high compositional variation; **l**, **l**: low sequencing depth and low compositional variation. 31

2.3 (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. 34

2.4 Inferred networks from each method with edges filtered by selection probability and ranked by edge weight, in comparison with the 91 interactions reported in literature. In each network, darker blue implies stronger (larger in absolute value) edge weight. 39

3.1 Normalized Manhattan similarity between the two estimated inverse covariance matrices with true and false references. Solid blue: Inv-Comp-gLASSO; dashed red: Inv-gLASSO. **h**, **h**: high sequencing depth, high compositional variation; **h**, **l**: high sequencing depth, low compositional variation; **l**, **h**: low sequencing depth, high compositional variation; **l**, **l**: low sequencing depth, low compositional variation. 57

3.2 Jaccard index between the two networks with true and false references. Solid blue: Inv-Comp-gLASSO; dashed red: Inv-gLASSO. **h**, **h**: high sequencing depth, high compositional variation; **h**, **l**: high sequencing depth, low compositional variation; **l**, **h**: low sequencing depth, high compositional variation; **l**, **l**: low sequencing depth, low compositional variation. . 60

Figure Page

3.3 Normalized Hamming similarity between the two networks with true and false references. Solid blue: Inv-Comp-gLASSO; dashed red: Inv-gLASSO. **h**,**h**: high sequencing depth, high compositional variation; **h**,**l**: high sequencing depth, low compositional variation; **l**,**h**: low sequencing depth, high compositional variation; **l**,**l**: low sequencing depth, low compositional variation. 61

3.4 ROC curves for Inv-Comp-gLASSO and Inv-gLASSO with true and false references. Long-dashed blue: Inv-Comp-gLASSO with true reference; dashed red: Inv-Comp-gLASSO with false reference; dashed-dotted purple: Inv-Comp-gLASSO with true reference; dotted black: Inv-Comp-gLASSO with false reference. **h**,**h**: high sequencing depth and high compositional variation; **h**,**l**: high sequencing depth and low compositional variation; **l**,**h**: low sequencing depth and high compositional variation; **l**,**l**: low sequencing depth and low compositional variation. 63

3.5 Normalized Manhattan similarity between the two estimated inverse covariance matrices with the the two choices of reference, Acrosphaera and Collosphaera, from the TARA data. Solid blue: Inv-Comp-gLASSO; dashed red: Inv-gLASSO. 66

3.6 Jaccard index and normalized Hamming similarity between the two estimated networks with the the two choices of reference, Acrosphaera and Collosphaera, from the TARA data. Solid blue: Inv-Comp-gLASSO; dashed red: Inv-gLASSO. 67

LIST OF TABLES

Table Page

2.1 For the hub genera, their rank in the degree distributions (in descending order) from the literature, compositional graphical lasso (Comp-gLASSO), graphical lasso (gLASSO) and neighborhood selection (MB). The numbers in the parentheses are the corresponding degrees of the genera. . . . 37

2.2 Top 15 genus pairs that were commonly predicted by compositional graphical lasso, graphical lasso, and neighborhood selection, and were not in the literature list. All genus pairs have selection probability 1, and are thus ranked by the sum of the absolute values of estimated partial correlations from the three methods. 41

Microbial Network Recovery by Compositional Graphical Lasso Under Additive Log-Ratio Transformation

1 Introduction

1.1 Challenges in Recovering Microbial Interactions

Microbiome, which is the collection of micro-organisms in an ecological system, is of great research interest in the science community. The advancement of the high-throughput sequencing technologies such as 16S rRNA profiling has enabled researchers to analyze the microbial compositions in uncultivated samples. By coupling the high-throughput sequencing procedures with bioinformatic and data analytic approaches, scientists have begun to disentangle the composition, diversity, and function of microbiomes, such as in the Human Microbiome Project (Huttenhower et al., 2012), the TARA Oceans Project (Sunagawa et al., 2015), and the Earth Microbiome Project (Thompson et al., 2017). Recent research has shown that microbiome could play an important role in influencing its host or living environment. For instance, it is found that intervening the gut microbiota of African turquoise killifish could result in delay of their aging process (Smith et al., 2017). Additionally, free-living microbes, such as those associated with soil and water, play central roles in modulating their environmental conditions (Clarholm, 1985; Paerl and Pinckney, 1996; Cotner and Biddanda, 2002; Bardgett et al., 2008; Wieder et al., 2013).

One common goal in microbiome data analysis is to understand how microbes interact with each other. The interactions may be beneficial or adverse, and are frequently important to the community and the environment (Faust and Raes, 2012). A profound understanding of the underlying mechanisms of those interactions as well as the means that disturb these interactions, is critical to the advancement of technologies in regulating microbiomes towards a favorable state, e.g., probiotic administration and microbiome transplantations (Sonnenburg and Fischbach, 2011; Nicholson et al., 2012). However, the nature of microbiome data imposed by the technicalities of the sequencing procedures has imposed various challenges in recovering the microbial interactions. We present these challenges as follows.

1.1.1 Compositionality

Microbiome data entangle with the "compositionality", which is a technicality imposed by the sequencing procedures. For instance, in 16S rRNA profiling, a specific sequence of marker genes which serves as the identifier of the Operational Taxonomic Units (OTU's, the surrogate of bacteria species) is sequenced and then counted as the proxy of the abundances of the corresponding OTU's in a sample. The total count of sequences in a sample, also known as the "sequencing depth", is pre-determined on the sequencing instrument and is usually not on the same scale from sample to sample. This implies that the counts for each OTU in a sample carry only the information about their relative abundances instead of their absolute abundance. We refer to this type of data as "compositional count data" to indicate both the compositional and discrete nature of microbiome data.

Interaction analysis disregarding the compositionality in the data, e.g., marginal correlation analysis of the OTUs' relative abundances, could result in spurious correlations as pointed out by Pearson (1897). A common approach to addressing this problem is to conduct a log-ratio transformation before analyzing the data (Aitchison, 1986). In particular, Aitchison proposed two types of log-ratio transformations: the centered log-ratio (CLR) transformation and the additive log-ratio (ALR) transformation. The CLR transformation centralizes the logarithms of the relative abundances, which results in a rank-deficient covariance matrix of the transformed vector and thus an undefined inverse covariance matrix. The ALR transformation takes the logarithm of the ratio of the relative abundances of the remaining entries against a preselected entry as the reference, which results in a transformed vector with one dimension lower. However, this transformed vector has a full-rank covariance matrix and thus a well-defined inverse covariance matrix. Currently, there are more interaction analysis methods based on the CLR transformation than the ALR transformation because the consequence regarding the different choices of the reference to the interaction analysis is still unclear.

1.1.2 Finite and Heterogeneous Sequencing Depth

One unique heterogeneity between studies or between samples in microbiome data is the variation of sequencing depths. As discussed above, sequencing depth, the total count of sequences generated across all taxa for a biological sample, is an experimental technicality and often varies considerably across samples in a microbiome sequencing experiment. The observed relative abundance of a taxon in a sample serves as an unbiased estimator of its true relative abundance, however, the variance of such an estimator depends on the sample-specific sequencing depth. For example, two equal relative abundances of an OTU in two samples can have unequal variances due to the different sequencing depths between the two samples. This problem of unequal variances is called “heteroscedasticity”.

Many interaction analyses simply treat the observed relative abundances as the true relative abundances without considering the estimation uncertainty and heteroscedasticity. An alternative approach is to normalize the count data before applying any analysis, with the most popular normalization method called “rarefaction”. By rarefying the data, this normalization method randomly subsamples the sequences without replacement until all sequencing depths equal to a “rarefaction level” after subsampling. The rarefaction level is usually set to the minimum sequencing depth in all samples. However, rarefaction is inefficient as it drops valuable data and introduces additional variation from subsampling. Neither ignoring heteroscedasticity nor trying to mitigate this problem by rarefaction works well in downstream analyses such as differential abundance analysis, where they both inflate the false positive rate (McMurdie and Holmes, 2014). In this dissertation, we will show that the estimation uncertainty in relative abundances and heteroscedasticity also affects the construction of microbial networks, which is indeed one of the major motivations of this work.

1.1.3 High-dimensionality

In addition to compositionality and heteroscedasticity, the microbiome data are often high-dimensional in nature. With the resolution at the OTU level, it is likely that the number of OTU’s is far more than the number of samples. In a typical microbiome dataset with decent sequencing depths, we have observed the number of OTU's in tens of thousands while the number of samples ranges from hundreds to thousands. The high-dimensionality feature imposes additional challenges on the interaction analysis of microbiome data. For example, the sample covariance (correlation) matrix computed from the abundances of the OTU's is rank-deficient, no matter how the abundances are normalized or transformed beforehand. In the high-dimensional setting, the classical estimators (such as the maximum likelihood estimator) of the covariance matrix or the inverse covariance matrix underperform more recently developed estimators such as those relying on regularization methods (Meinshausen et al., 2006; Yuan and Lin, 2007; Banerjee et al., 2008; Friedman et al., 2008).

1.2 Previous Work on Microbial Interaction Networks

Existing work that focused on recovering the interaction network among microbes can be categorized into two main types: the ones that model the marginal correlations, and the others that model the conditional dependences. We are going to review both types of methods as follows.

1.2.1 Marginal Correlation Networks

After one obtains the abundance data for the microbial species, marginal correlation analysis could be used to infer the interactions among microbes (Faust and Raes, 2012). However, naive approaches that ignore the compositionality of the data, e.g., simply calculating the Pearson's or Spearman's correlations from the abundances or relative abundances would lead to spurious correlations (Pearson, 1897). More specifically, two taxa that are indeed independent with each other could seem to be negatively correlated under this approach, due to the compositional constraint.

Over the years, several methods have been developed to address the compositionality issue in the construction of correlation networks for microbiome data, such as SparCC (Friedman and Alm, 2012) , CCLasso (Fang et al., 2015), and REBECCA (Ban et al., 2015). All these methods aimed to construct a covariance (correlation) matrix or network of the unknown absolute abundances (the positive, unconstrained true abundances of taxa). In particular, SparCC iteratively selects the most highly correlated pairs of taxa until reaching certain correlation threshold or violating certain sparsity assumption. On the other hand, both CCLasso and REBECCA have an explicit objective function and obtain the estimate of the covariance matrix by solving the optimization problem. It is noteworthy that the high-dimensionality in microbiome data is tackled by imposing a sparsity constraint in the above three methods, with a correlation threshold for SparCC and an L_1-norm penalty for CCLasso and REBECCA.

However, none of the three methods took the finite and heterogeneous sequencing depth into account. All three methods applied a log-ratio transformation of the count data as their first step, thus ignored the estimation uncertainty and heteroscedasticity issues as mentioned in Section 1.1.2.

1.2.2 Conditional Dependence Networks

The edges in a marginal correlation network include indirect microbial interactions. For example, two taxa that do not have a direct dependence between each other could have a non-zero correlation because they both directly interact with a third taxon. Conditional dependence networks, which only capture the conditional dependence between two taxa conditioning on all the other taxa, are therefore often more desirable.

Many existing methods for conditional dependence networks are based on graphical models. For example, in a Gaussian graphical model with the multivariate normal distribution, two variables are conditionally independent given all the others if and only if the entry corresponding to those two variables in the inverse covariance matrix is zero. Therefore, a conditional dependence network can be inferred through an estimation problem of the inverse covariance matrix, a.k.a., concentration matrix or precision matrix. However, in the high-dimensional setting, i.e. when the number of variables is larger than the number of samples, the maximum likelihood estimate of the covariance matrix is degenerate, and its inverse is thus unavailable. Under the assumption that the true inverse covariance matrix is sparse, the graphical lasso (Yuan and Lin, 2007; Banerjee et al., 2008; Friedman et al., 2008) and neighborhood selection (Meinshausen et al., 2006) solve this problem by imposing an L_1-norm penalty. Although graphical lasso and neighborhood selection are useful, they are however not customized for microbiome data analysis—the microbiome data are compositional counts that are discrete along with a fixed total in each sample, thus clearly not following the multivariate Gaussian distribution.

Several methods have been proposed to infer microbial conditional dependence networks based on Gaussian graphical models, such as SPIEC-EASI (Kurtz et al., 2015), gCoda (Fang et al., 2017), CD-trace (Yuan et al., 2019), and SPRING (Yoon et al., 2019). In order to transform the discrete counts to continuous variables and to remove the compositionality constraint, all these methods take the CLR transformation as their first step. However, the CLR transformation suffers from an undefined inverse covariance matrix of the transformed data (see Section 1.1.1) . All methods resolve this problem by imposing a sparsity assumption on the inverse covariance matrix and adding an L_1-norm penalty to their objective functions. Unfortunately, all these methods ignore the finite and heterogeneous sequencing depth as they simply treat the observed relative abundances as the truth. As mentioned in Section 1.1.2, this is essentially ignoring the uncertainty and heteroscedasticity in the estimation of relative abundances, which could impact downstream analysis.

To the best of our knowledge, MLDM (Yang et al., 2016) is one of the few methods that address the finite and heterogeneous sequencing depth issue by explicitly modeling the compositional counts in a Dirichlet-multinomial distribution. However, their hierarchical model is three-leveled with numerous ancillary parameters (including the regression coefficients of the environmental covariates, which are not always needed for the purpose of estimating a microbial interaction network), rendering an algorithm lack of computational efficiency and scalability. Besides, theoretical property on the convergence of the algorithm is not available, raising additional concerns on its applicability.

Although most of the existing methods are based on graphical models, FlashWeave (Tackmann et al., 2019), completely built on hypothesis testing, has also been proposed to infer conditional dependences for microbiome data. For each taxon, FlashWeave first selects its "neighborhood", such as taxa that are marginally correlated with it, as the "candidates of conditionally dependent taxa". Then, it conducts a series of conditional testing similar to the "forward selection" procedure in model selection. Unlike forward selection that only tests the significance of a new variable conditioning on the whole previously selected set, it tests the significance of a new variable conditioning on all the subsets of the previously selected variables, leading to a high computational complexity. To reduce the runtime, it has to limit the size of the candidate set and rely on an early stopping rule, introducing considerable false positives and false negatives in its result. Moreover, it also takes the CLR transformation as its first step and thus ignores the finite and heterogeneous sequencing depth issue like most other methods that we have reviewed so far.

1.3 Our Contribution

In Chapter 2 of this dissertation, we first propose a novel method called "compositional graphical lasso" that estimates the conditional dependence network from the microbiome data through a hierarchical model called logistic normal multinomial model, accounting for all three features as reviewed in Section 1.1: compositionality, finite and heterogeneous sequencing depth, and high-dimensionality. The method is based on the ALR transformation, in which a reference taxon needs to be selected as a first step. One natural concern would be whether the different choices of the reference taxon affects the subsequent network analysis. To address this concern, we establish in Chapter 3 the reference-invariance property of a subnetwork corresponding to the non-candidate-reference taxa. Moreover, we propose a reference-invariant version of compositional graphical lasso by a simple modification on the penalty function in the compositional graphical lasso objective function. It is noteworthy that similar modifications can be potentially applied to other methods for conditional dependence networks that are based on the ALR-transformed data. Hopefully, this work can stimulate the research interest on network analysis methods based on the ALR transformation of the microbial abundance data. Some relevant further discussion can be found in Chapter 4 of the dissertation.

2 Microbial Network Recovery by Compositional Graphical Lasso

2.1 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.

2.2 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 (Yuan and Lin, 2007; Banerjee et al., 2008; Friedman et al., 2008) and neighborhood selection (Meinshausen et al., 2006) are two commonly used methods to estimate sparse inverse covariance matrix for high-dimensional data under the Gaussian graphical models . However, both the counting and the compositional features of the microbiome abundance data have violated the multivariate normality assumption.

SPIEC-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. SPIEC-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 SPIEC-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 an oceanic microbiome data set.

2.3 Compositional Graphical Lasso

2.3.1 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:

$$\mathbf{x}_i|\mathbf{p}_i \sim \text{Multinomial}(M_i; p_{i,1}, \ldots, p_{i,K+1}), \tag{2.1}$$

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:

$$z_{i,k} = \log\left(\frac{p_{i,k}}{p_{i,K+1}}\right),\ i = 1, \ldots, n,\ k = 1, \ldots, K. \tag{2.2}$$

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

$$\mathbf{z}_1, \ldots, \mathbf{z}_n \overset{iid}{\sim} N(\mu, \Sigma), \tag{2.3}$$

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

The above model in (2.1)–(2.3) 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 Ω in (2.3). 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 Σ that does not require the logistic normal assumption. By contrast, we aim to find an estimator of Ω directly based on the logistic normal multinomial model.

2.3.2 Objective Function

From the logistic normal multinomial model in (2.1)–(2.3), 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 $\mathbf{x}_i$,

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

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

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

Given the multivariate normal parameters μ and Ω, 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

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

Since our objective is to find a sparse estimator of the inverse covariance matrix Ω, we further maximizes the above maximum posterior distribution over μ and Ω with a L_1 penalty on Ω, or equivalently,

$$\begin{aligned}&\min_{\mu,\Omega} -\frac{1}{n}\log[f_{\mu,\Omega}(\hat{\mathbf{z}}_1,\ldots,\hat{\mathbf{z}}_n|\mathbf{x}_1,\ldots,\mathbf{x}_n)] + \lambda\|\Omega\|_1\\ &= \min_{\mu,\Omega}\min_{\mathbf{z}_1,\ldots,\mathbf{z}_n} -\frac{1}{n}\log[f_{\mu,\Omega}(\mathbf{z}_1,\ldots,\mathbf{z}_n|\mathbf{x}_1,\ldots,\mathbf{x}_n)] + \lambda\|\Omega\|_1. \qquad (2.4)\end{aligned}$$

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

$$\begin{aligned}\ell(\mathbf{z}_1,\ldots,\mathbf{z}_n,\mu,\Omega) = & -\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=1}^{K} x_{i,k}z_{i,k} - M_i \log\left(\sum_{k=1}^{K} e^{z_{i,k}}+1\right)\right] \\ & -\frac{1}{2}\log[\det(\Omega)] + \frac{1}{2n}\sum_{i=1}^{n}(\mathbf{z}_i-\mu)'\Omega(\mathbf{z}_i-\mu)+\lambda\|\Omega\|_1. \qquad (2.5)\end{aligned}$$

In other words, $\mathbf{z}_1,\ldots,\mathbf{z}_n$, μ, and Ω are all treated as unknown parameters for minimization in the objective function (2.5).

The objective function (2.5) was introduced as profiling out the parameters $(\mathbf{z}_1,\ldots,\mathbf{z}_n)$ and then minimizing the posterior distribution over μ and Ω. However, there is a Bayesian interpretation as well. Notice that we have set up a logistic normal multinomial model as in (2.1)–(2.3). If we treat μ and Ω as random and impose a hyperprior distribution on them as

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

then, (2.5) becomes the negative logarithm of the posterior distribution of $\mathbf{z}_1,\ldots,\mathbf{z}_n,\mu,\Omega$ given $\mathbf{x}_1,\ldots,\mathbf{x}_n$. Therefore, minimizing (2.5) is equivalent to finding the maximum a posteriori (MAP) estimator of $\mathbf{z}_1,\ldots,\mathbf{z}_n,\mu,\Omega$ given the data $\mathbf{x}_1,\ldots,\mathbf{x}_n$.

2.3.3 Computational Algorithm

The objective function (2.5) includes naturally three sets of parameters $(\mathbf{z}_1,\ldots,\mathbf{z}_n)$, μ, and Ω, 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 $(\mathbf{z}_1^{(0)},\ldots,\mathbf{z}_n^{(0)})$, $\mu^{(0)}$, and $\Omega^{(0)}$, a block coordinate algorithm repeats the following steps cyclically for iteration $t = 0, 1, 2, \ldots$ until the algorithm converges.

1. Given $\mu^{(t)}$ and $\Omega^{(t)}$, find $(\mathbf{z}_1^{(t+1)}, \ldots, \mathbf{z}_n^{(t+1)})$ that maximizes (2.5).
2. Given $(\mathbf{z}_1^{(t+1)}, \ldots, \mathbf{z}_n^{(t+1)})$ and $\Omega^{(t)}$, find $\mu^{(t+1)}$ that maximizes (2.5).
3. Given $(\mathbf{z}_1^{(t+1)}, \ldots, \mathbf{z}_n^{(t+1)})$ and $\mu^{(t+1)}$, find $\Omega^{(t+1)}$ that maximizes (2.5).

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 $\mu^{(0)}$, notice that we have a closed form minimizer of μ for (2.5) given the values of $(\mathbf{z}_1, \ldots, \mathbf{z}_n)$, which is $\mu = \bar{\mathbf{z}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i$. Therefore, we set the initial value as $\mu^{(0)} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i^{(0)}$. Finally, for the initial value $\Omega^{(0)}$, 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 $\mu^{(t)}$ and $\Omega^{(t)}$, minimizing the objective function (2.5) 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 $i = 1, \ldots, n$:

$$\ell_i^{(t)}(\mathbf{z}_i) = \frac{1}{2}(\mathbf{z}_i - \mu^{(t)})'\Omega^{(t)}(\mathbf{z}_i - \mu^{(t)}) - \left[\sum_{k=1}^{K} x_{i,k} z_{i,k} - M_i \log\left(\sum_{k=1}^{K} e^{z_{i,k}} + 1\right)\right]. \qquad (2.6)$$

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 μ 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 $\mu^{(t+1)} = \bar{\mathbf{z}}^{(t+1)}$, the objective function for Ω can be simplified as

$$\begin{aligned}\ell^{(t)}(\Omega) &= -\frac{1}{2}\log[\det(\Omega)] + \frac{1}{2n}\sum_{i=1}^{n}(\mathbf{z}_i^{(t+1)} - \mu^{(t+1)})'\Omega(\mathbf{z}_i^{(t+1)} - \mu^{(t+1)}) + \lambda\|\Omega\|_1, \\ &= -\frac{1}{2}\log[\det(\Omega)] + \frac{1}{2}\mathrm{tr}(\mathbf{S}^{(t+1)}\Omega) + \lambda\|\Omega\|_1, \qquad (2.7)\end{aligned}$$

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 (2.7) becomes a graphical lasso problem (Yuan and Lin, 2007; Banerjee et al., 2008; Friedman et al., 2008). It is well known that the graphical lasso objective function is a convex function in Ω (Banerjee et al., 2008) 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.

2.3.4 Theoretical Convergence

Unfortunately, the objective function (2.5) is not necessarily a convex function jointly in $(\mathbf{z}_1,\dots,\mathbf{z}_n)$, μ, and Ω. However, we have shown that it is actually convex in each subset of its parameters (see Section 2.3.3). 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 $(\mathbf{z}_1,\dots,\mathbf{z}_n)$, μ, and Ω as three blocks and optimize for each block iteratively. In addition, as in Tseng (2001), the objective function (2.5) can be regarded as the sum of two parts, the first of which is an inseparable but differentiable function as

$$\ell_0(\mathbf{z}_1,\dots,\mathbf{z}_n,\mu,\Omega) = \frac{1}{2n}\sum_{i=1}^{n}(\mathbf{z}_i-\mu)'\Omega(\mathbf{z}_i-\mu), \tag{2.8}$$

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

$$\ell_1(\mathbf{z}_1,\ldots,\mathbf{z}_n) = -\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=1}^{K} x_{i,k}z_{i,k} - M_i \log\left(\sum_{k=1}^{K} e^{z_{i,k}}+1\right)\right], \tag{2.9}$$

$$\ell_2(\Omega) = -\frac{1}{2}\log[\det(\Omega)] + \lambda\|\Omega\|_1. \tag{2.10}$$

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 $\mathscr{A} \subset \mathbb{R}^n$ is a real vector $\mathbf{a} \in \mathbb{R}^n$ such that for every $\delta > 0$, there exists a point $\mathbf{x}$ in $\mathscr{A} \setminus \{\mathbf{a}\}$ such that $\|\mathbf{x}-\mathbf{a}\|_2 < \delta$. Obviously, any limit point of the set $\mathscr{A}$ is a cluster point.

Furthermore, define a cluster point of the compositional graphical lasso algorithm to be a cluster point of the set $\{(\mathbf{z}_1^{(t)},\ldots,\mathbf{z}_n^{(t)},\mu^{(t)},\Omega^{(t)}) : t = 0,1,2,\ldots\}$, 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. *Consider a bounded and open parameter space of $(\mathbf{z}_1,\ldots,\mathbf{z}_n,\mu,\Omega)$. Then, any cluster point of the compositional graphical lasso algorithm in this parameter space is a stationary point of the objective function (2.5).*

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, $\ell_0(\mathbf{z}_1,\ldots,\mathbf{z}_n,\mu,\Omega)$ in (2.8) is regular at each point in its domain. This is true because $\mathrm{dom}(\ell_0)$ is open and ℓ_0 is differentiable and all its partial derivatives exists.

Second, the level set $\{(\mathbf{z}_1,\dots,\mathbf{z}_n,\mu,\Omega) : \ell(\mathbf{z}_1,\dots,\mathbf{z}_n,\mu,\Omega) \le \ell(\mathbf{z}_1^{(0)},\dots,\mathbf{z}_n^{(0)},\mu^{(0)},\Omega^{(0)})\}$ is compact and that ℓ in (2.5) 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 (2.5) is bounded below.

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

where

$$\hat{\Omega} = \arg\min_{\Omega} -\frac{1}{2}\log[\det(\hat{\Omega})] + \frac{1}{2n}\sum_{i=1}^{n}(\mathbf{z}_i-\mu)'\hat{\Omega}(\mathbf{z}_i-\mu) + \lambda\|\hat{\Omega}\|_1,$$

and the last inequality follows because $\hat{\Omega}$ is unique and has positive eigenvalues bounded above by $\frac{K}{\lambda}$ (Banerjee et al., 2008). Therefore, the level set can be written as $\{(\mathbf{z}_1,\dots,\mathbf{z}_n,\mu,\Omega) : c_1 \le \ell(\mathbf{z}_1,\dots,\mathbf{z}_n,\mu,\Omega) \le c_2\}$ for some constant c_1 and c_2, thus is close as it is a preimage of a close set under a continuous function. Furthermore, since we consider a bounded parameter space of $(\mathbf{z}_1,\dots,\mathbf{z}_n,\mu,\Omega)$, the level set is also compact.

Third, $\ell(\mathbf{z}_1,\dots,\mathbf{z}_n,\mu,\Omega)$ has at most one minimum in its second block of parameters, i.e., μ. This is true given that the Hessian matrix for μ is Ω 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., $\{\ell(\mathbf{z}_1^{(t)},\ldots,\mathbf{z}_n^{(t)},\mu^{(t)},\Omega^{(t)}) : t = 0,1,2,\ldots\}$ 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., $\{(\mathbf{z}_1^{(t)},\ldots,\mathbf{z}_n^{(t)},\mu^{(t)},\Omega^{(t)}) : t = 0,1,2,\ldots\}$. Instead, Theorem 1 provides some theoretical guarantees about the convergence of the minimizers, which states that any cluster point of the algorithm in the considered parameter space 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.

2.3.5 Tuning Parameter Selection

There is a large body of literature on the selection of 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 apply StARS to select the tuning parameter λ in our objective function (2.5). 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 Ω, 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.

Starting from a large penalty which corresponds to the empty network, the instability of networks increases as λ decreases. StARS stops and selects the tuning parameter to be the minimum value of λ's 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 densest among networks with a total instability less than a threshold β without violate some sparsity assumption. The selected network is the "densest on the sparse side", as it starts with the empty network and stops when the instability first across the threshold. When λ becomes really small and eventually estimates dense and even fully-connected networks (though StARS stops way before this happens in reality), the instability could decrease and drop below the threshold again, since every edge is always selected with small enough λ.

2.4 Simulation Study

2.4.1 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 and graphical lasso.

Given our goal is to estimate the true network, i.e., Ω in (2.3), we consider three types of precision matrix $\Omega = (\omega_{kl})_{1 \le k,l \le K}$, 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 Ω 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 (2.1). We simulate M_i from two uniform distributions, $\text{Unif}(20K, 40K)$ and $\text{Unif}(100K, 200K)$, and call the two settings low and high sequencing depth, respectively. The second factor is the variation included in the logistic normal distribution (2.3). Although we consider three types of precision matrices, we consider an additional factor by multiplying a positive constant c to Ω so that the true precision matrix is $c\Omega$. 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 (2.1)–(2.3). We first simulate $\mathbf{z}_i \sim N(\mu, \Sigma)$ independently for $i = 1, \ldots, n$; then, we perform the inverse log-ratio transformation (also know as the softmax transformation, the inverse transformation of (2.2)) 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 Ω. For neighborhood selection and graphical lasso, we first obtain an estimate of $\mathbf{z}_1, \dots, \mathbf{z}_n$ from the multinomial distribution as

$$\tilde{z}_{i,k} = \log\left(\frac{x_{i,k}}{x_{i,K+1}}\right),\ i = 1, \dots, n,\ k = 1, \dots, K, \tag{2.11}$$

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

To compare the performance of the three methods in terms of network recovery, all three methods are applied with a sequence of 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.

2.4.2 Results

Figure 2.1 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 neighborhood 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 (2.11) 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

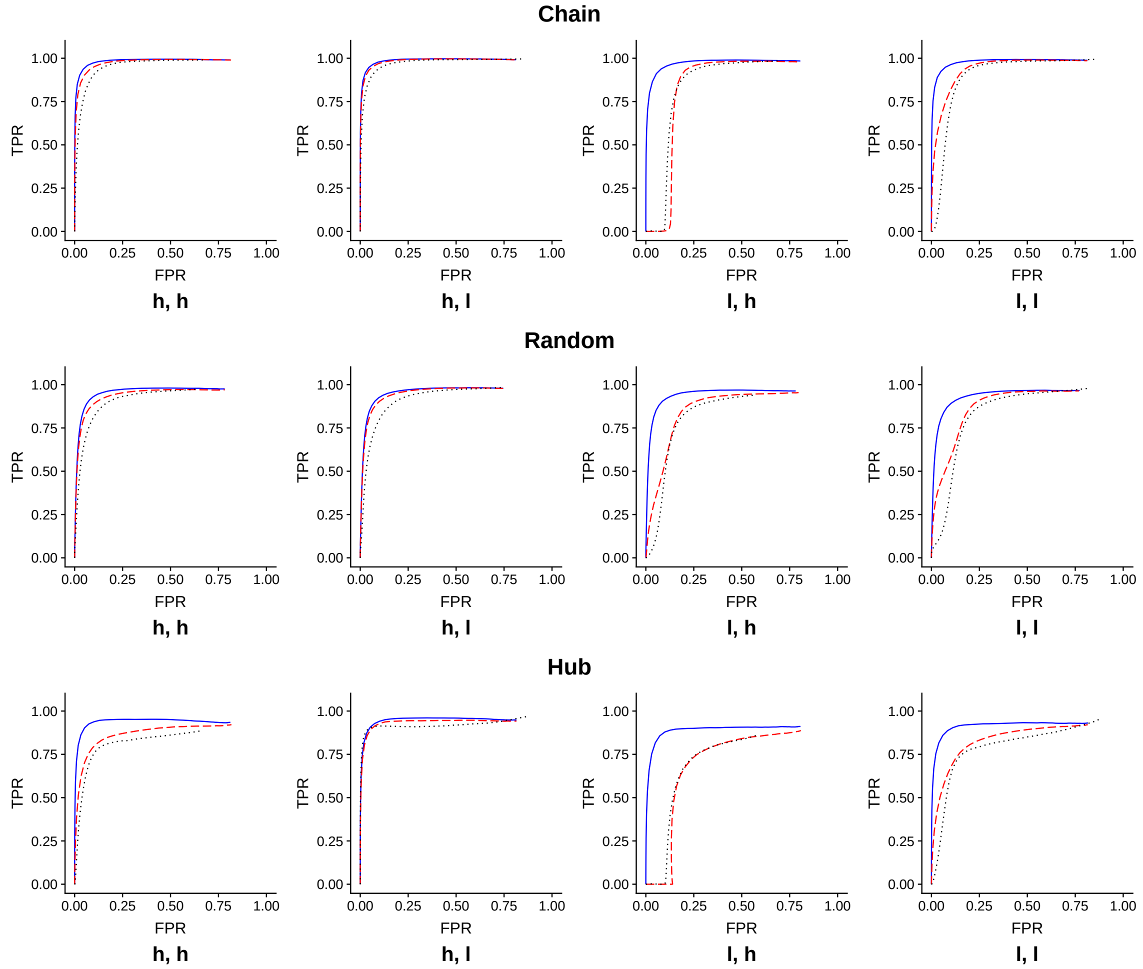

Figure 2.1: ROC curves for compositional graphical lasso (Comp-gLASSO), graphical lasso (gLASSO) and neighborhood selection (MB). Solid blue: Comp-gLASSO; dashed red: gLASSO; dotted black: MB. **h**, **h**: high sequencing depth and high compositional variation; **h**, **l**: high sequencing depth and low compositional variation; **l**, **h**: low sequencing depth and high compositional variation; **l**, **l**: low sequencing depth and low compositional variation.

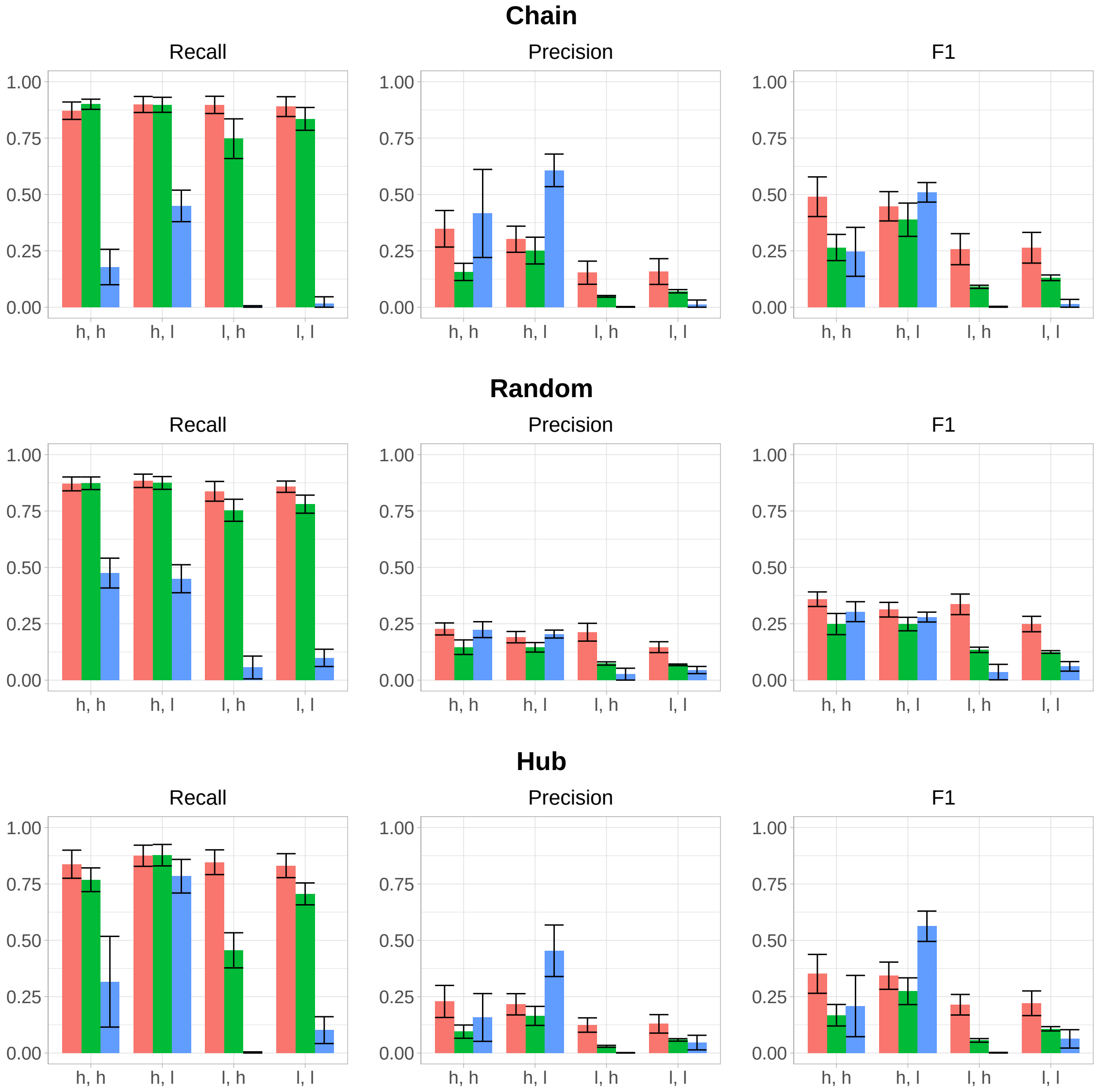

Figure 2.2: Recall, precision and F1 score for the network selected by StARS for compositional graphical lasso (Comp-gLASSO), graphical lasso (gLASSO) and neighborhood selection (MB). Red (left): Comp-gLASSO; green (middle): gLASSO; blue (right): MB. **h**, **h**: high sequencing depth and high compositional variation; **h**, **l**: high sequencing depth and low compositional variation; **l**, **h**: low sequencing depth and high compositional variation; **l**, **l**: low sequencing depth and low compositional variation.

multinomial component in the model, it is also not surprising to see that their performances are deteriorated by a high compositional variation.

Figure 2.2 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, 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 (within the instability 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 precision 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 or comparable precision, recall, and F1 score than the ones from graphical lasso in all settings. 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.

2.5 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 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 the genus Acrosphaera, which has the largest average relative abundance among those genera not involved in the gold-standard list, 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.

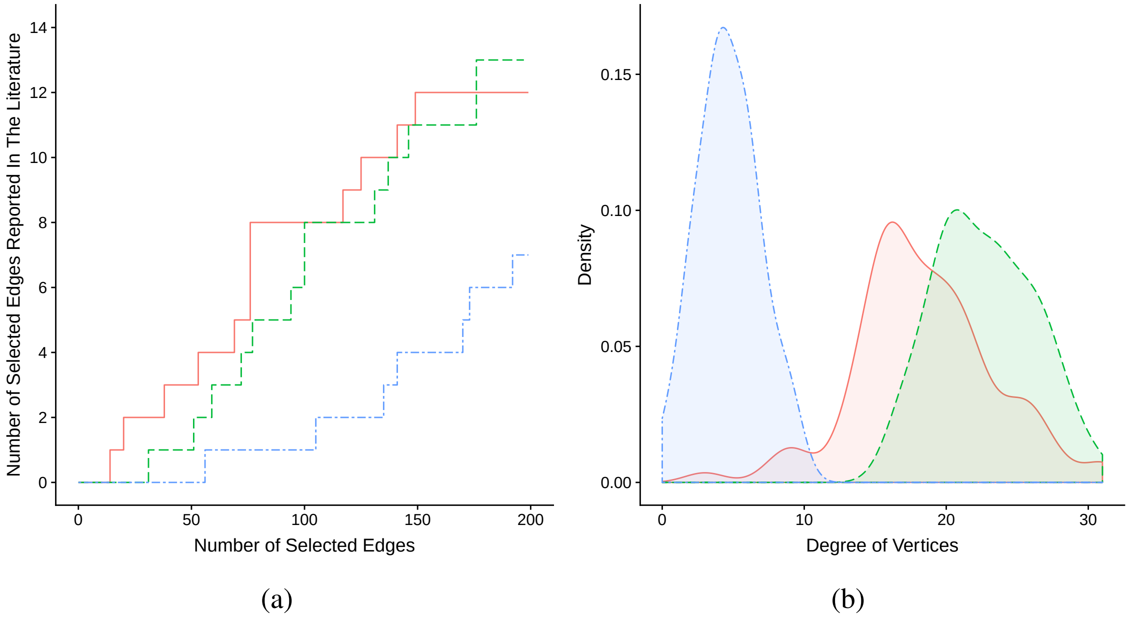

Figure 2.3: (a): Number of identified literature interactions versus number of edges of the estimated network from the TARA dataset. (b): The degree distribution of vertices from the networks selected by StARS. Solid red: compositional graphical lasso; dashed green: graphical lasso; dashed dotted blue: neighbor hood selection.

First, to compare the three methods in terms of their ability to reconstruct the literature interactions, we apply each of them with 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 2.3a.

From Figure 2.3a, 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 2.3b. It looks like the degree distribution of the network from neighborhood selection is highly right-skewed, the one from graphical lasso is quite left-skewed, and 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 the ascending order, which also reflects that the densities of the final 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. Although the literature-reported interactions are 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 degrees of the hub genera from the three networks directly, but it is reasonable to compare the rank of those degrees within each degree distributions. The method that generates 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 the literature is shown in Table 2.1, followed by the corresponding ranks of those genera in the degree distributions from each of the three methods and their corresponding degrees in the three estimated networks in the parentheses. We can see that compositional graphical lasso generates lower ranks than graphical lasso for all 7 genera, while neighborhood selection generates 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 methods.

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:

	Literature	Comp-gLASSO	gLASSO	MB
Amoebophrya	1 (21)	31 (19)	47 (21)	2 (9)
Blastodinium	2 (12)	13 (23)	26 (24)	30 (5)
Parvilucifera	2 (12)	46 (17)	67 (19)	58 (3)
Syndinium	4 (7)	14 (23)	27 (24)	19 (6)
Vampyrophrya	4 (7)	34 (19)	60 (20)	6 (8)
Phaeocystis	6 (6)	1 (31)	3 (29)	17 (6)
Pirsonia	7 (5)	64 (15)	68 (19)	33 (5)

Table 2.1: For the hub genera, their rank in the degree distributions (in descending order) from the literature, compositional graphical lasso (Comp-gLASSO), graphical lasso (gLASSO) and neighborhood selection (MB). The numbers in the parentheses are the corresponding degrees of the genera.

(a) selection probability, 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 further ranked by edge weight. For the networks from all three methods, darker blue implies higher magnitude in the absolute value of partial correlation.

We can see from Figure 2.4 that, though still different, the networks estimated by the three algorithms have apparent similarity in the predicted edges and their edge weights, e.g., the genus pairs "Centropages - Thalassicolla" and "Acanthometra - Hexaconus" are in dark blue for all three methods. On the other hand, there are 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 selection probability and larger weight might suggest promising new eukaryotic interactions that deserve biological validations. There are actually 39 common edges from the top 100 edges from the three estimated networks. We further ranked them by (a) the sum of selection probabilities from the three networks and (b) the sum of edge weights estimated from the three methods, and provided the list of the top 15 genus pairs in the Appendix for the interested readers.

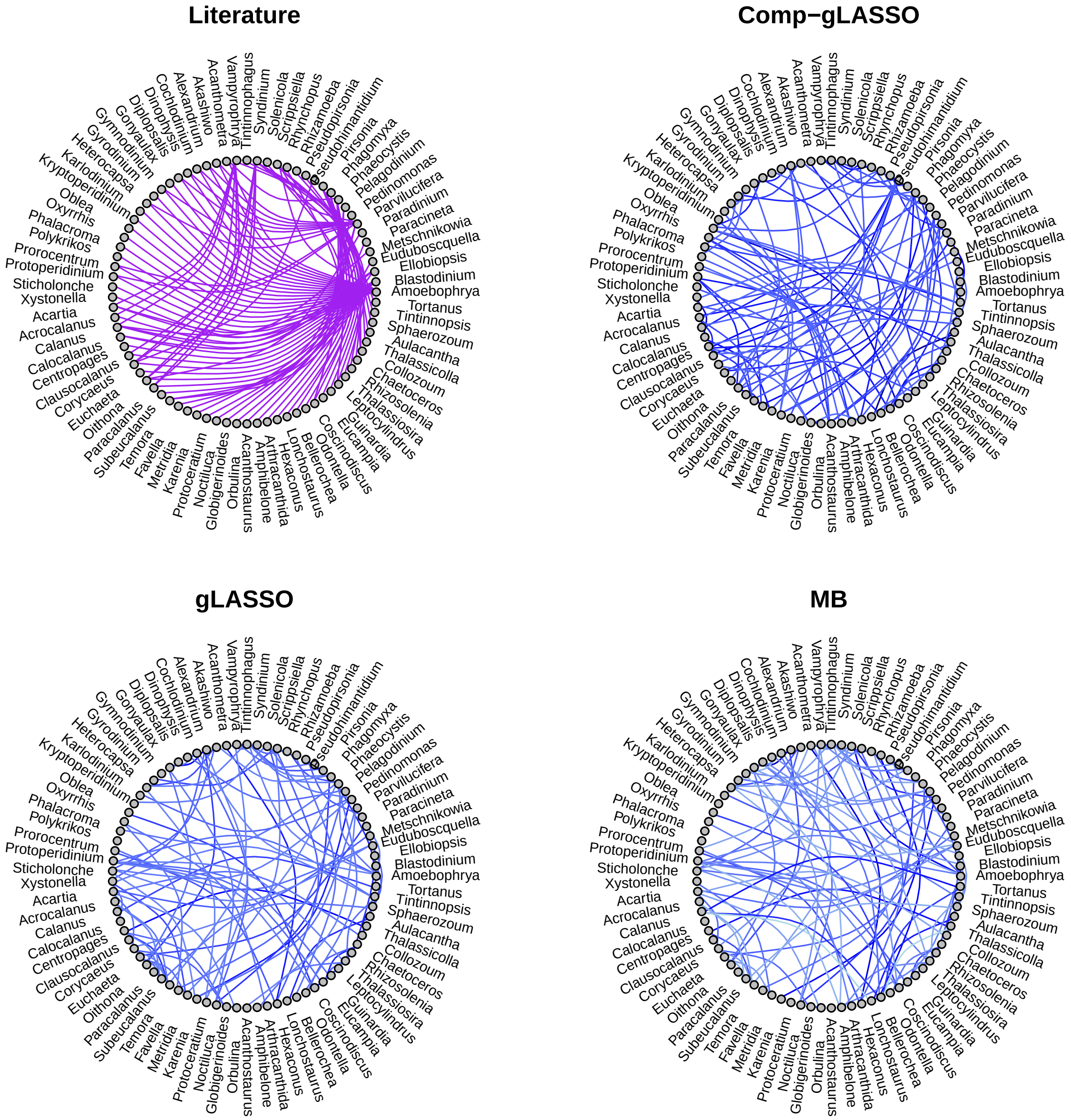

Figure 2.4: Inferred networks from each method with edges filtered by selection probability and ranked by edge weight, in comparison with the 91 interactions reported in literature. In each network, darker blue implies stronger (larger in absolute value) edge weight.

2.6 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 the analyses of synthetic and real data will come up soon.

2.7 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 2.2: Top 15 genus pairs that were commonly predicted by compositional graphical lasso, graphical lasso, and neighborhood selection, and were not in the literature list. All genus pairs have selection probability 1, and are thus ranked by the sum of the absolute values of estimated partial correlations from the three methods.

3 Reference-Invariance Property of Inverse Covariance Matrix Estimation Under Additive Log-Ratio Transformation and Its Application to Microbial Network Recovery

3.1 Abstract

The interactions between microbial taxa in microbiome data has been under great research interest in the science community. In particular, several methods such as SPIEC-EASI, gCoda, and CD-trace have been proposed to model the conditional dependency between microbial taxa, in order to eliminate the detection of spurious correlations. However, all those methods are built upon the central log-ratio (CLR) transformation, which results in a degenerate covariance matrix and thus an undefined inverse covariance matrix as the estimation of the underlying network. Jiang et al. (2020) and Tian et al. (2020) proposed bias-corrected graphical lasso and compositional graphical lasso based on the additive log-ratio (ALR) transformation, which first selects a reference taxon and then computes the log ratios of the abundances of all the other taxa with respect to that of the reference. One concern of the ALR transformation would be the invariance of the estimated network with respect to the choice of reference. In this paper, we first establish the reference-invariance property of a subnetwork of interest based on the ALR transformed data. Then, we propose a reference-invariant version of the compositional graphical lasso by modifying the penalty in its objective function, penalizing only the invariant subnetwork. We validate

the reference-invariance property of the proposed method under a variety of simulation scenarios as well as through the application to an oceanic microbiome data set.

3.2 Introduction

Microbiome, which is the collection of micro-organisms in an ecological system, is of great research interest in the science community and has been shown to play an important role in influencing its host or living environment. For instance, it is found that intervening the gut microbiota of African turquoise killifish could result in delay of their aging process (Smith et al., 2017). The advancement of the high-throughput sequencing technologies such as 16S rRNA profiling that replicates a specific sequence of marker genes which is counted and serves as the proxy of the abundance of the Operational Taxonomic Units (OTU's, the surrogate of bacteria species) in a sample, has enabled researchers to analyze the microbial compositions in uncultivated samples.

One common goal in microbiome data analysis is to understand how microbes interact with each other. However, the nature of microbiome data determined by the technicalities of the sequencing procedures has imposed various challenges in recovering the microbial interactions. Firstly, the data is composed of discrete counts. Secondly, microbiome data entangles with the "compositionality", which is a technicality imposed by the sequencing procedures. For instance, in 16S rRNA profiling, the "sequencing depth", i.e. the total count of sequences in a sample, is pre-determined on the sequencing instrument, and is usually not on the same scale from sample to sample. This implies that the counts for each OTU in a sample carry only the information about their relative abundances instead of their absolute abundances. Thirdly, the microbiome data possesses “high-dimensionality” in nature. With the resolution at the OTU level, it is likely that the number of OTU’s is far more than the number of samples in a biological experiment.

After one obtains the abundance data for the microbial species, marginal correlation analysis could be used to infer the interactions among microbes (Faust and Raes, 2012). Over the years, several methods have been developed to address the compositionality issue in the construction of correlation networks for microbiome data, such as SparCC (Friedman and Alm, 2012) , CCLasso (Fang et al., 2015), and REBECCA (Ban et al., 2015). All of those methods aimed to construct a covariance (correlation) matrix or network of the unknown absolute abundances (the positive, unconstrained true abundance of taxa). The high-dimensionality in microbiome data is tackled by imposing a sparsity constraint in the above three methods, with a correlation threshold for SparCC and an L_1-norm penalty for CCLasso and REBECCA.

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. SPIEC-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 (CLR) transformation on the observed counts (Aitchison, 1986), and then apply graphical lasso (Yuan and Lin, 2007; Banerjee et al., 2008; Friedman et al., 2008) to find the inverse covariance matrix of the transformed data. More recently, gCoda and CD-trace were developed to improve SPIEC-EASI by accounting for the compositionality property of microbiome data (Fang et al., 2017; Yuan et al., 2019), both of which have been shown to possess better performance in terms of recovering the sparse microbial network than SPIEC-EASI.

It is worth noting that SPIEC-EASI, gCoda, and CD-trace are all built upon the CLR transformation of the observed counts. Meanwhile, Jiang et al. (2020) and Tian et al. (2020) proposed bias-corrected graphical lasso and compositional graphical lasso based on the additive log-ratio (ALR) transformed data. In ALR transformation, one needs to select a reference taxon and compute the log relative abundance of all other taxa with respect to the reference. One of the major concerns for the ALR transformation is the robustness or invariance of the proposed method with respect to the choice of the reference taxon, which is not well studied in the literature.

In this paper, we first establish the reference-invariance property of estimating the sparse microbial network based on the ALR transformed data. It shows that a submatrix of the inverse covariance matrix that correspond to the non-candidate-of-reference taxa is invariant with respect to the choice of the reference. Then, we propose a reference-invariant version of the compositional graphical lasso by modifying the penalty in its objective function, which only penalizes the invariant submatrix mentioned above. Additionally, we illustrate the reference-invariance property of the proposed method under a variety of simulation scenarios and also demonstrate its applicability and advantages by applying it to an oceanic microbiome data set.

3.3 Methodology

3.3.1 Reference-Invariance Property

Let $\mathbf{p} = (p_1, \ldots, p_{K+2})'$ denote a vector of compositional probabilities satisfying that $p_1 + \cdots + p_{K+2} = 1$. The additive log-ratio (ALR) transformation picks an entry of this vector as the reference and transforms the compositional vector using log ratios of each entry to the reference. Without loss of generality, suppose we pick the last entry as the reference, then the ALR transformed vector becomes

$$\mathbf{z} = \left[\log\left(\frac{p_1}{p_{K+2}}\right), \ldots, \log\left(\frac{p_K}{p_{K+2}}\right), \log\left(\frac{p_{K+1}}{p_{K+2}}\right)\right]'.$$

The transformed vector $\mathbf{z}$ is often assumed to follow a multivariate continuous distribution with a mean vector μ and a covariance matrix Σ. For example, $\mathbf{z} \sim N(\mu, \Sigma)$. Denote further the inverse covariance matrix $\Omega = \Sigma^{-1}$.

Similarly, if we pick another entry as the reference, we can define another ALR-transformed vector. For simplicity of illustration, suppose we choose the second last entry to be the reference and consider the following ALR transformation

$$\mathbf{z}_p = \left[\log\left(\frac{p_1}{p_{K+1}}\right), \ldots, \log\left(\frac{p_K}{p_{K+1}}\right), \log\left(\frac{p_{K+2}}{p_{K+1}}\right)\right]',$$

where the subscript p denotes the "permuted" version of $\mathbf{z}$. Similarly, define the mean vector of $\mathbf{z}_p$ by μ_p, the covariance matrix by Σ_p, and the inverse covariance matrix by $\Omega_p = \Sigma_p^{-1}$.

A simple derivation implies that $\mathbf{z}_p$ is a linear transformation of $\mathbf{z}$ as $\mathbf{z}_p = \mathbf{Q}_p\mathbf{z}$, where

$$\mathbf{Q}_p = \begin{pmatrix} \mathbf{I}_{K+1} & -\mathbf{1} \\ \mathbf{0}' & -1 \end{pmatrix}$$

with $\mathbf{I}_{K+1}$ denoting the identity matrix, and $\mathbf{0}$ and $-\mathbf{1}$ denoting the column vectors with all 0's and all -1's, respectively. It follows that $\mu_p = \mathbf{Q}_p\mu$, $\Sigma_p = \mathbf{Q}_p\Sigma\mathbf{Q}_p'$, and $\Omega_p = (\mathbf{Q}_p')^{-1}\Omega\mathbf{Q}_p^{-1}$. It is also worth noting that $\mathbf{Q}_p$ is an involutory matrix, i.e., $\mathbf{Q}_p^{-1} = \mathbf{Q}_p$.

The following theorem states the reference-invariance property of the inverse covariance matrix Ω under the ALR transformation.

Theorem 2. $\Omega_{1:K,1:K} = \Omega_{p,1:K,1:K}$*, i.e. the* $K \times K$ *upper-left sub-matrix of the inverse covariance matrix of the* ALR *transformed vector is invariant with respect to the choices of the* $(K+2)$*-th entry or the* $(K+1)$*-th entry as the reference.*

Theorem 2 regards the reference-invariance property of the true value of the inverse covariance matrix Ω. It can also be extended to a class of estimators of Ω. Suppose we have i.i.d. observations of the compositional vectors $\mathbf{p}_1,\ldots,\mathbf{p}_n$, and consequently, their ALR transformed counterparts $\mathbf{z}_1,\ldots,\mathbf{z}_n$. Then, we can construct an estimator of Σ, denoted by $\hat{\Sigma}$, based on the i.i.d. observations $\mathbf{z}_1,\ldots,\mathbf{z}_n$. Furthermore, we can construct an estimator of Ω, denoted by $\hat{\Omega}$, by taking its inverse or generalized inverse. The following corollary presents the reference-invariance property for a class of such estimators.

Corollary 1. *Suppose* $\hat{\Sigma}_p = \mathbf{Q}_p\hat{\Sigma}\mathbf{Q}_p'$ *and both* $\hat{\Sigma}_p$ *and* $\hat{\Sigma}$ *are invertible. Let* $\hat{\Omega} = \hat{\Sigma}^{-1}$ *and* $\hat{\Omega}_p = \hat{\Sigma}_p^{-1}$ *be their inverse matrices. Then,* $\hat{\Omega}_{1:K,1:K} = \hat{\Omega}_{p,1:K,1:K}$*, i.e. the* $K \times K$ *upper-left sub-matrix of the estimated inverse covariance matrix of the* ALR *transformed vector is* *invariant with respect to the choices of the $(K+2)$-entry or the $(K+1)$-th entry as the reference.*

The above results imply an important property for the additive log-ratio transformation in the compositional data analysis. It can be extended to a more general situation as follows. In general, suppose we have selected a set of entries $\mathbf{p}_{\mathscr{R}}$ as "candidate references" in a compositional vector $\mathbf{p}$ and write $\mathbf{p} = (\mathbf{p}'_{\mathscr{R}^c}, \mathbf{p}'_{\mathscr{R}})'$. Then, for any ALR transformed vector $\mathbf{z}$ based on a reference in the set of candidate references $\mathbf{p}_{\mathscr{R}}$, the $|\mathscr{R}^c| \times |\mathscr{R}^c|$ upper-left sub-matrix of the (estimated) inverse covariance matrix of $\mathbf{z}$ is invariant with respect to the choice of the reference.

In the following subsections, we will incorporate the reference-invariance property into the estimation of a sparse inverse covariance matrix for compositional count data, such as the OTU abundance data in microbiome research.

3.3.2 Logistic Normal Multinomial Model

Consider an OTU abundance data set with n independent samples, each of which composes observed counts of $K+2$ taxa, denoted by $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,K+2})'$ for the i-th sample, $i = 1, \ldots, n$. Due to the compositional property of the data, the sum of all counts for each sample i is a fixed number, denoted by M_i. Naturally, a multinomial distribution is imposed on the observed counts as

$$\mathbf{x}_i | \mathbf{p}_i \sim \text{Multinomial}(M_i, \mathbf{p}_i), \tag{3.1}$$

where $\mathbf{p}_i = (p_{i,1}, \ldots, p_{i,K+2})'$ are the multinomial probabilities with $\sum_{k=1}^{K+2} p_{i,k} = 1$.

In addition, we choose one taxon, without loss of generality, the $(K+2)$-th taxon as the reference and then apply the ALR transformation (Aitchison, 1986) on the multinomial probabilities as follows

$$\mathbf{z}_i = \left[\log\left(\frac{p_{i,1}}{p_{i,K+2}}\right), \ldots, \log\left(\frac{p_{i,K}}{p_{i,K+2}}\right), \log\left(\frac{p_{i,K+1}}{p_{i,K+2}}\right)\right]', \; i = 1, \ldots, n. \tag{3.2}$$

Further assume that $\mathbf{z}_i$'s follow an i.i.d. multivariate normal distribution

$$\mathbf{z}_i \overset{iid}{\sim} N(\mu, \Sigma), \; i = 1, \ldots, n, \tag{3.3}$$

where μ is the mean, Σ is the covariance matrix, and $\Omega = \Sigma^{-1}$ is the inverse covariance matrix. The above model in (3.1)–(3.3) is called a logistic normal multinomial model and has been applied to analyze the microbiome abundance data (Xia et al., 2013).

Tian et al. (2020) proposed a method called compositional graphical lasso that aims to find a sparse estimator of the inverse covariance matrix Ω, in which the following objective function is minimized

$$\begin{aligned}\ell(\mathbf{z}_1, \ldots, \mathbf{z}_n, \mu, \Omega) = & -\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{x}'_{i,-(K+2)}\mathbf{z}_i - M_i \log\{\mathbf{1}'\exp(\mathbf{z}_i) + 1\}\right] \\ & -\frac{1}{2}\log[\det(\Omega)] + \frac{1}{2n}\sum_{i=1}^{n}(\mathbf{z}_i - \mu)'\Omega(\mathbf{z}_i - \mu) + \lambda\|\Omega\|_1,\end{aligned} \tag{3.4}$$

where $\mathbf{x}_{i,-(K+2)} = (x_{i,1}, \ldots, x_{i,K+1})'$ and $\mathbf{1} = (1, \ldots, 1)'$. The above objective function has two parts: The first term in (3.4) is the negative log-likelihood of the multinomial distribution in (3.1) and the remaining terms are the regular objective function of graphical lasso for the multivariate normal distribution in (3.3) regarding $\mathbf{z}_1, \ldots, \mathbf{z}_n$ as known quantities.

3.3.3 Reference-Invariant Objective Function

Similar to 3.3.1, if we choose another taxon, for simplicity of illustration, the $(K+1)$-th taxon as the reference, then the ALR transformation in (3.2) becomes

$$\mathbf{z}_{i,p} = \left[\log\left(\frac{p_{i,1}}{p_{i,K+1}}\right), \ldots, \log\left(\frac{p_{i,K}}{p_{i,K+1}}\right), \log\left(\frac{p_{i,K+2}}{p_{i,K+1}}\right)\right]'.$$

As in Sections 3.3.1, $\mathbf{z}_{i,p} = \mathbf{Q}_p \mathbf{z}_i$. Therefore, $\mathbf{z}_{i,p} \overset{iid}{\sim} N(\mu_p, \Sigma_p)$, $i = 1, \ldots, n$, where $\mu_p = \mathbf{Q}_p \mu$, $\Sigma_p = \mathbf{Q}_p \Sigma \mathbf{Q}_p'$, and $\Omega_p = (\mathbf{Q}_p')^{-1} \Omega \mathbf{Q}_p^{-1}$. The reference-invariance property in 3.3.1 implies that $\Omega_{1:K,1:K} = \Omega_{p,1:K,1:K}$.

The different choice of the reference also leads to a different objective function for the compositional graphical lasso method (Comp-gLASSO) as follows

$$\begin{aligned}\ell_p(\mathbf{z}_{1,p}, \ldots, \mathbf{z}_{n,p}, \mu_p, \Omega_p) = & -\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{x}_{i,-(K+1)}' \mathbf{z}_{i,p} - M_i \log\{\mathbf{1}' \exp(\mathbf{z}_{i,p}) + 1\}\right] \\ & -\frac{1}{2}\log[\det(\Omega_p)] + \frac{1}{2n}\sum_{i=1}^{n}(\mathbf{z}_{i,p} - \mu_p)' \Omega_p (\mathbf{z}_{i,p} - \mu_p) + \lambda \|\Omega_p\|_1,\end{aligned} \tag{3.5}$$

where $\mathbf{x}_{i,-(K+1)} = (x_{i,1}, \ldots, x_{i,K}, x_{i,K+2})'$ and $\mathbf{1} = (1, \ldots, 1)'$. Comparing (3.4) and (3.5), their first terms are the same as they are both equal to the negative log-likelihood of the multinomial distribution: $-\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K+2} x_{i,k} \log p_{i,k}$. In addition, from Aitchison (1986), $\det(\Omega) = \det(\Omega_p)$ and $\sum_{i=1}^{n}(\mathbf{z}_i - \mu)'\Omega(\mathbf{z}_i - \mu) = \sum_{i=1}^{n}(\mathbf{z}_{i,p} - \mu_p)'\Omega_p(\mathbf{z}_{i,p} - \mu_p)$ as known properties of the ALR transformation. However, the L_1 penalties in (3.4) and (3.5) are different because Ω is not necessarily equal to Ω_p. The reference-invariance property only implies that $\Omega_{1:K,1:K} = \Omega_{p,1:K,1:K}$.

Motivated by the reference-invariance property, we can impose the L_1 penalties only on the invariant entries of Ω instead of all entries of Ω as in (3.4), which leads to

$$\begin{aligned}\ell_{inv}(\mathbf{z}_1,\ldots,\mathbf{z}_n,\mu,\Omega) = &-\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{x}'_{i,-(K+2)}\mathbf{z}_i - M_i\log\{\mathbf{1}'\exp(\mathbf{z}_i)+1\}\right] \\ &-\frac{1}{2}\log[\det(\Omega)]+\frac{1}{2n}\sum_{i=1}^{n}(\mathbf{z}_i-\mu)'\Omega(\mathbf{z}_i-\mu)+\lambda\|\Omega_{1:K,1:K}\|_1.\end{aligned} \tag{3.6}$$

With the previous arguments, we showed that $\ell_{\text{inv}}(\mathbf{z}_1,\ldots,\mathbf{z}_n,\mu,\Omega)$ is reference-invariant; in other words, the objective function ℓ_{inv} stays the same regardless of whether the $(K+1)$-th or the $(K+2)$-th taxa is selected as the reference. This is summarized in the following theorem.

Theorem 3. *If* $\mathbf{z}_{i,p} = \mathbf{Q}_p\mathbf{z}_i$ *for* $i = 1,\ldots,n$, $\mu_p = \mathbf{Q}_p\mu$, *and* $\Omega_p = (\mathbf{Q}'_p)^{-1}\Omega\mathbf{Q}_p^{-1}$, *then* $\ell_{inv}(\mathbf{z}_1,\ldots,\mathbf{z}_n,\mu,\Omega) = \ell_{inv,p}(\mathbf{z}_{1,p},\ldots,\mathbf{z}_{n,p},\mu_p,\Omega_p)$.

We call $\ell_{inv}(\mathbf{z}_1,\ldots,\mathbf{z}_n,\mu,\Omega)$ in (3.6) the reference-invariant compositional graphical lasso objective function. To obtain a sparse estimator of Ω, we minimize the objective function $\ell_{inv}(\mathbf{z}_1,\ldots,\mathbf{z}_n,\mu,\Omega)$ with respect to $\mathbf{z}_1,\ldots,\mathbf{z}_n$, μ, and Ω. We call this estimation approach the reference-invariant compositional graphical lasso (Inv-Comp-gLASSO) method.

In general, suppose we have selected a set of taxa $\mathbf{x}_{\mathscr{R}}$ as "candidate references" and write $\mathbf{x} = (\mathbf{x}'_{\mathscr{R}^c}, \mathbf{x}'_{\mathscr{R}})'$. Then, the reference-invariant objective function becomes

$$\begin{aligned}\ell_{\text{inv}}(\mathbf{z}_1, \ldots, \mathbf{z}_n, \mu, \Omega) = & -\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{x}'_{i,\mathscr{R}^c}\mathbf{z}_i - M_i \log\{\mathbf{1}'\exp(\mathbf{z}_i) + 1\}\right] \\ & -\frac{1}{2}\log[\det(\Omega)] + \frac{1}{2n}\sum_{i=1}^{n}(\mathbf{z}_i - \mu)'\Omega(\mathbf{z}_i - \mu) + \lambda\|\Omega_{\mathscr{R}^c,\mathscr{R}^c}\|_1.\end{aligned} \tag{3.7}$$

In other words, (3.7) is invariant regardless of which reference is selected in the set of candidate references, and so is the invariant part of its minimizer.

It is noteworthy that the trick we played in defining the reference-invariant version of Comp-gLASSO is to revise the penalty term from the regular lasso penalty on the whole inverse covariance matrix to that only on the invariant part of the inverse covariance matrix. Using the same trick, we can define the reference-invariant version of other methods such as reference-invariant graphical lasso (Inv-gLASSO). The objective function of Inv-gLASSO is defined as follows when $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are observed instead of $\mathbf{x}_1, \ldots, \mathbf{x}_n$:

$$\ell_{\text{inv}}(\mu, \Omega) = -\frac{1}{2}\log[\det(\Omega)] + \frac{1}{2n}\sum_{i=1}^{n}(\mathbf{z}_i - \mu)'\Omega(\mathbf{z}_i - \mu) + \lambda\|\Omega_{\mathscr{R}^c,\mathscr{R}^c}\|_1. \tag{3.8}$$

The objective function (3.7) includes naturally three sets of parameters $(\mathbf{z}_1, \ldots, \mathbf{z}_n)$, μ, and Ω, 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 $(\mathbf{z}_1^{(0)}, \ldots, \mathbf{z}_n^{(0)})$, $\mu^{(0)}$, and $\Omega^{(0)}$, a block coordinate algorithm repeats the following steps cyclically for iteration $t = 0, 1, 2, \ldots$ until the algorithm converges.

1. Given $\mu^{(t)}$ and $\Omega^{(t)}$, find $(\mathbf{z}_1^{(t+1)}, \ldots, \mathbf{z}_n^{(t+1)})$ that maximizes (3.7).

2. Given $(\mathbf{z}_1^{(t+1)}, \ldots, \mathbf{z}_n^{(t+1)})$ and $\Omega^{(t)}$, find $\mu^{(t+1)}$ that maximizes (3.7).

3. Given $(\mathbf{z}_1^{(t+1)}, \ldots, \mathbf{z}_n^{(t+1)})$ and $\mu^{(t+1)}$, find $\Omega^{(t+1)}$ that maximizes (3.7).

Except for the mild modification on the objective function, the details of the algorithm is essentially the same as the one described in Tian et al. (2020).

3.4 Simulation Study

3.4.1 Settings

To illustrate the reference-invariance property under the aforementioned framework, we conduct a simulation study and evaluated the performance of Inv-Comp-gLASSO as well as Inv-gLASSO.

Following Tian et al. (2020), we generate three types of inverse covariance matrices $\Omega = (\omega_{kl})_{1 \leq k,l \leq K+1}$ as follows:

1. **Chain**: $\omega_{kk} = 1.5$, $\omega_{kl} = 0.5$ if $|k-l| = 1$, and $\Omega_{kl} = 0$ if $|k-l| > 1$. Every node is connected to the adjacent node(s), and therefore the degree is 2 for all but two nodes.

2. **Random**: $\omega_{kl} = 1$ with probability $3/(K+1)$ for $k \neq l$. Every two nodes are connected with a fixed probability, and the expected degree is the same for all nodes.

3. **Hub**: Nodes are randomly partitioned into $\lceil (K+1)/20 \rceil$ groups, and there's one "hub node" in each group. For the other nodes in the group, they are only connected

to the hub node but not each other. There's no connection among groups. The degree of connectedness is much higher for the hub nodes , and is 1 for the rest of the nodes.

In the simulations, we also vary two other factors that play a crucial role in the performances of the methods:

1. **Sequencing depth**. M_i's are simulated from $\text{Uniform}(20K, 40K)$ or $\text{Uniform}(100K, 200K)$, denoted by "low" and "high" sequencing depth.

2. **Compositional variation**. For each aforementioned inverse covariance matrix Ω ("low" compositional variation), we also divide each of them by a factor of 5 to obtain another set of inverse covariance matrices, i.e., $\Omega/5$ ("high" compositional variation).

The data are simulated from the logistic normal multinomial distribution in (3.1)–(3.3). In detail, $\mathbf{z}_i \sim N(\mu, \Sigma)$ are first generated independently for $i = 1, \ldots, n$; then, the softmax transformation (the inverse of the ALR transformation) was applied to get the multinomial probabilities $\mathbf{p}_i$ with the $(K+2)$-th entry serving as the true reference; last, the multinomial random variables $\mathbf{x}_i$ were simulated from $Multinomial(M_i; \mathbf{p}_i)$, for $i = 1, \ldots, n$. We set $n = 100$ and $K = 49$ throughout the simulations.

The simulation results are based on 100 replicates of the simulated data. Both Inv-Comp-gLASSO and Inv-gLASSO are applied with two choices of reference, the $(K+1)$-th entry serving as the false reference and the $(K+2)$-th entry serving as the true reference. and only the reference-invariant sub-network $\Omega_{1:K,1:K}$ is used in the evaluations. For Inv-gLASSO, we estimate $\mathbf{p}_1, \ldots, \mathbf{p}_n$ with $\mathbf{x}_1/M_1, \ldots, \mathbf{x}_n/M_n$, and performe the ALR transformation to get the estimates of $\mathbf{z}_1, \ldots, \mathbf{z}_n$, which are denoted by $\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_n$. We then apply Inv-gLASSO to $\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_n$ directly to find the inverse covariance matrix Ω, which also serves as the starting value for Inv-Comp-gLASSO. For both methods, we implement them with a common sequence of 70 tuning parameters of λ.

We empirically validate the invariance property of Inv-Comp-gLASSO and Inv-gLASSO by comparing the estimators of the two sub-networks $\Omega_{1:K,1:K}$ resulted from choosing the true and false reference separately in each method, which have been shown to be theoretically invariant in Section 3.3. The comparison is assessed under four criteria as follows.

1. Normalized Manhattan Similarity

   For two matrices $\mathbf{A}$ and $\mathbf{B}$, we define the normalized Manhattan similarity (NMS) as

   $$\mathrm{NMS}(\mathbf{A}, \mathbf{B}) = 1 - \frac{\|\mathbf{A} - \mathbf{B}\|_1}{\|\mathbf{A}\|_1 + \|\mathbf{B}\|_1},$$

   where $\|\cdot\|_1$ represents the entrywise L_1 norm of a matrix. Note that $0 \leq \mathrm{NMS} \leq 1$ due to the non-negativity of norms and the triangle inequality.

2. Jaccard Index

   For two networks with the same nodes, denote their set of edges by $\mathscr{A}$ and $\mathscr{B}$. Then the Jaccard Index (Jaccard, 1901) is defined as follows:

   $$J(\mathscr{A}, \mathscr{B}) = \frac{|\mathscr{A} \cap \mathscr{B}|}{|\mathscr{A} \cup \mathscr{B}|}.$$

   Obviously, it also holds that $0 \leq J(\mathscr{A}, \mathscr{B}) \leq 1$.

3. Normalized Hamming Similarity

   In the context of network comparison, the normalized Hamming similarity for two adjacency matrices $\mathbf{A}$ and $\mathbf{B}$ with the same N nodes are defined as follows (Hamming, 1950)

$$H(\mathbf{A},\mathbf{B}) = 1 - \frac{\|\mathbf{A}-\mathbf{B}\|_1}{N(N-1)},$$

where $\|\cdot\|_1$ denotes the entrywise L_1 norm of a matrix. Since there are at most $N(N-1)$ 1's in an adjacency matrix with N nodes, this metric is also between 0 and 1.

4. ROC Curve

   The ROC curves on which true positive rate (TPR) and false positive rate (FPR) are plotted and also compared between the choices of true and false reference.

3.4.2 Results

3.4.2.1 Normalized Manhattan Similarity

The normalized Manhattan similarity serves as a direct measure of the similarity between the two estimated inverse covariance matrices with different choices of the reference. On a sequence of 70 tuning parameter λ's, we calculated the normalized Manhattan similarity between the two estimated inverse covariance matrices with the true and false references separately. Figure 3.1 shows the average normalized Manhattan similarity over 100 replicates along with standard error bars.

We can see from Figure 3.1 that, the normalized Manhattan similarity for Inv-gLASSO stays close to 1 in all settings, throughout the whole sequence of tuning parameters. On the other hand, there are some fluctuations in the same metric from Inv-Comp-gLASSO,

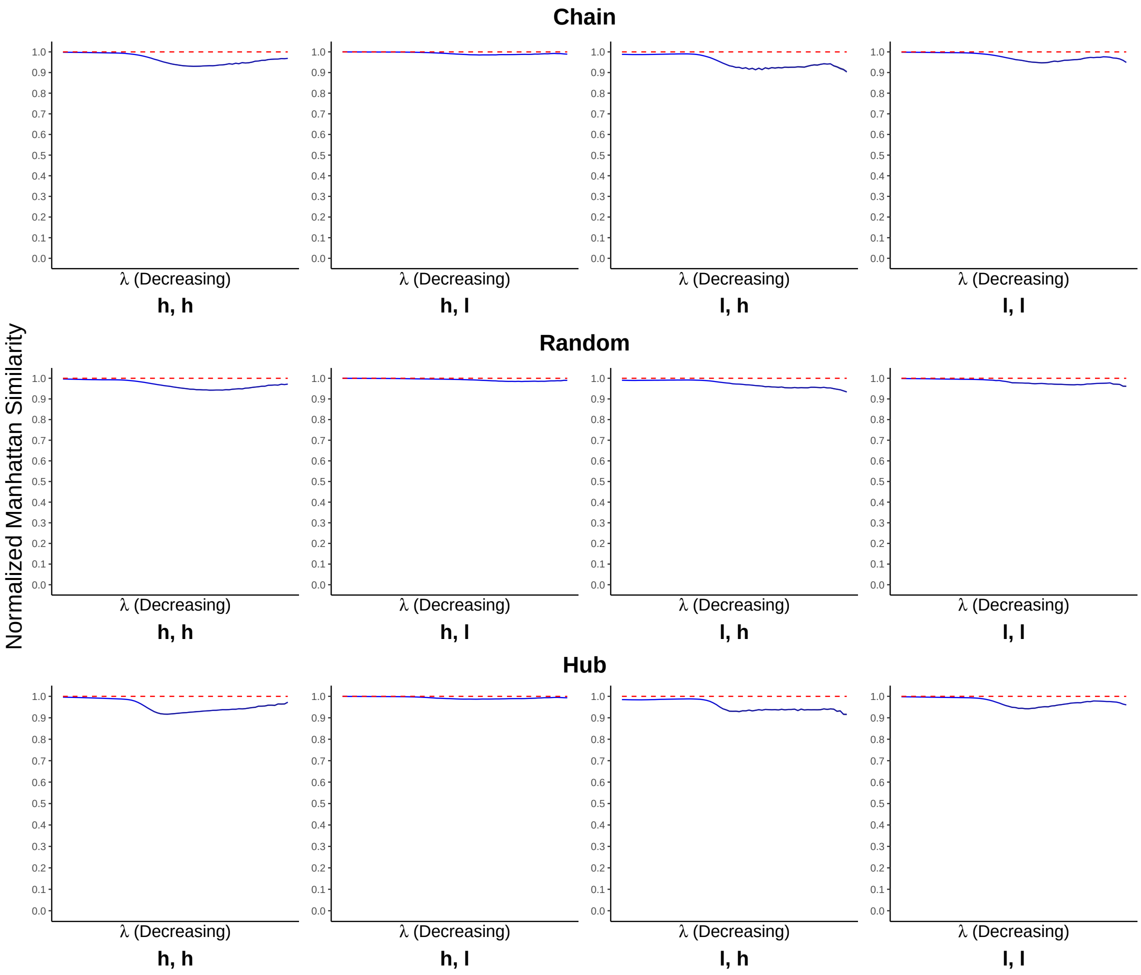

Figure 3.1: Normalized Manhattan similarity between the two estimated inverse covariance matrices with true and false references. Solid blue: Inv-Comp-gLASSO; dashed red: Inv-gLASSO. **h**, **h**: high sequencing depth, high compositional variation; **h**, **l**: high sequencing depth, low compositional variation; **l**, **h**: low sequencing depth, high compositional variation; **l**, **l**: low sequencing depth, low compositional variation.

although most values stay higher than 0.9. Empirically, the two estimated matrices are numerically identical for Inv-gLASSO and close for Inv-Comp-gLASSO. A potential reason why the invariance of Inv-gLASSO is numerically more evident than that of Inv-Comp-gLASSO is that Inv-gLASSO is a convex optimization while Inv-Comp-gLASSO is not necessarily convex (Tian et al., 2020). With different starting points (as we choose different references), Inv-Comp-gLASSO might result in different solutions as the algorithm is only guaranteed to converge to a stationary point. We refer to Tian et al. (2020) for more detailed discussion about the convexity and the convergence of the algorithm.

In addition, it is consistently observed that the normalized Manhattan similarity for Inv-Comp-gLASSO starts close to 1 when λ is very large and gradually decreases with some fluctuations as λ decreases. This is because the Inv-Comp-gLASSO objective function in (3.6) is solved by an iterative algorithm between graphical lasso to estimate (μ, Ω) and Newton-Raphson to estimate $\mathbf{z}_1, \ldots, \mathbf{z}_n$, which can lead to more numerical errors depending on the number of iterations. Furthermore, the algorithm is implemented with warm start for a sequence of decreasing λ's, i.e., the solution for the previous λ value is used as the starting point for the current λ value. With the accumulation of numerical errors, the numerical difference between the two estimated matrices becomes larger.

Among the simulation settings, we find the invariance property for Inv-Comp-gLASSO is most evidently supported by the numerical results in the "high sequencing depth, low compositional variation" setting, regardless of the network types. The normalized Manhattan similarity is very close to 1 for Inv-Comp-gLASSO throughout the sequence of tuning parameters. This is because the compositional probabilities $\mathbf{p}_i$'s and thus the $\mathbf{z}_i$'s are estimated accurately with high sequencing depth and low compositional variation in the first iteration of the Inv-Comp-gLASSO algorithm, which implies fewer iterations for the algorithm to converge and less numeric error accumulated during this process. On the other hand, the normalized Manhattan similarity is the lowest in the “low sequencing depth, high compositional variation” setting. It is due to a similar reason that it takes more iterations for the Inv-Comp-gLASSO algorithm to converge, accumulating more numerical errors. However, it is noteworthy that this is exactly the setting in which Inv-Comp-gLASSO has the most advantage over Inv-gLASSO in recovering the true network (see Section 3.4.2.3 for their ROC curves).

3.4.2.2 Jaccard Index and Normalized Hamming Similarity

Compared to normalized Manhattan similarity that measures directly the similarity between two inverse covariane matrices, both Jaccard index and normalized Hamming similarity measure the similarity between two networks represented by the matrices because they only compared the adjacency matrices or the edges of the two networks. Again, on a sequence of 70 tuning parameter λ’s, we computed the Jaccard index and the normalized Hamming similarity between the two networks with true and false references separately. Figures 3.2 and 3.3 plot the average Jaccard index and the normalized Hamming similarity over 100 replicates along with standard error bars.

The results of normalized Hamming similarity in Figure 3.3 have a fairly similar pattern to those of normalized Manhattan similarity in Figure 3.1 and thus can be similarly interpreted. We only focus on the results of Jaccard index in Figure 3.2 here. The Jaccard index stays close to 1 for Inv-gLASSO, implying the identity between the two networks.

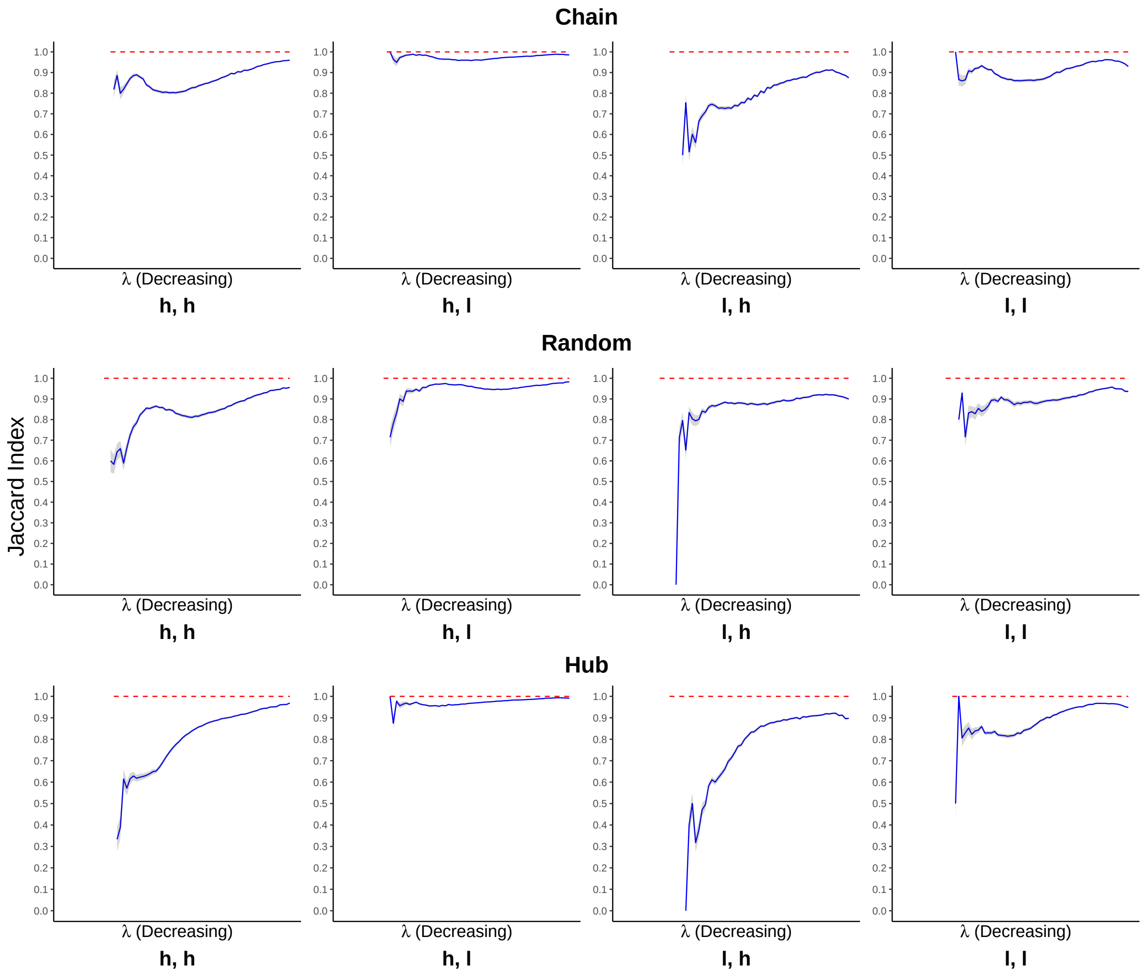

Figure 3.2: Jaccard index between the two networks with true and false references. Solid blue: Inv-Comp-gLASSO; dashed red: Inv-gLASSO. **h**, **h**: high sequencing depth, high compositional variation; **h**, **l**: high sequencing depth, low compositional variation; **l**, **h**: low sequencing depth, high compositional variation; **l**, **l**: low sequencing depth, low compositional variation.

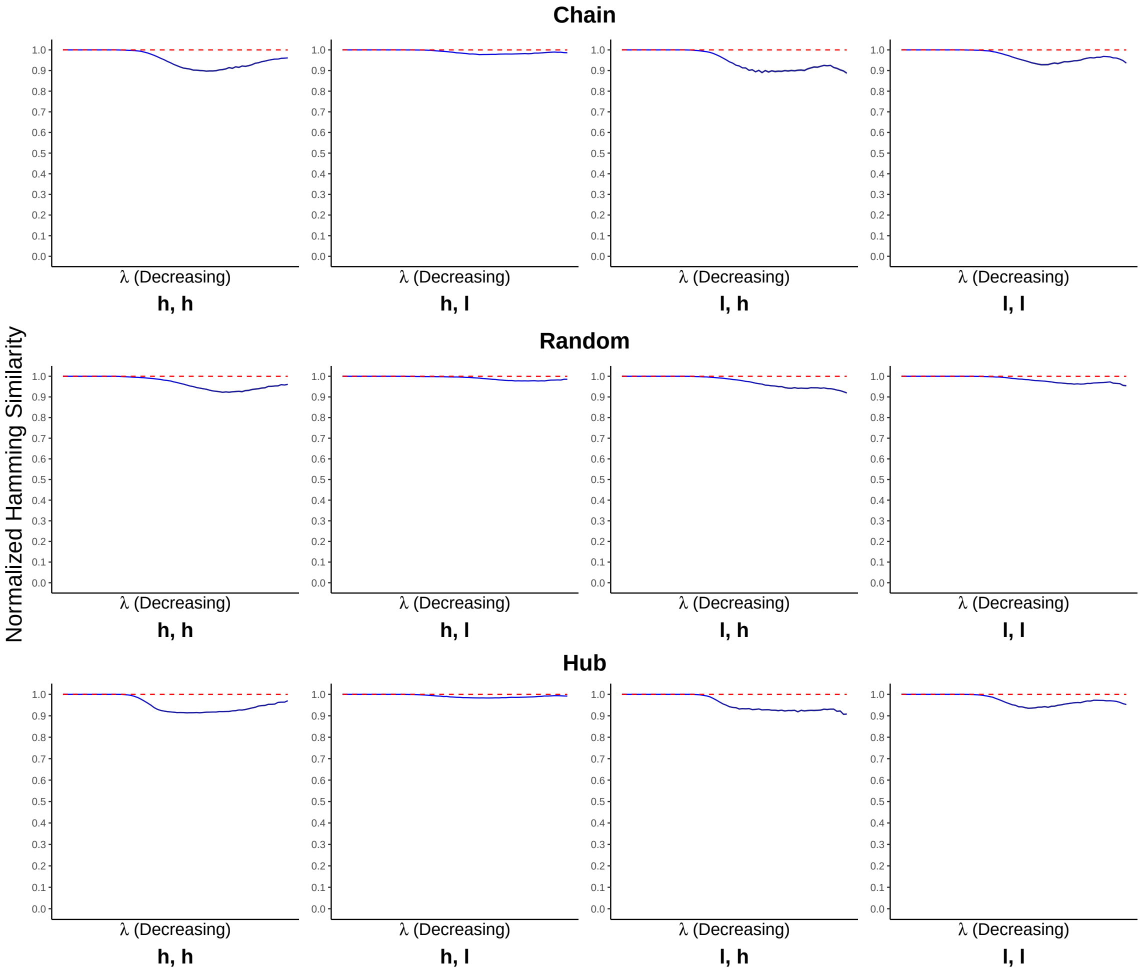

Figure 3.3: Normalized Hamming similarity between the two networks with true and false references. Solid blue: Inv-Comp-gLASSO; dashed red: Inv-gLASSO. **h**, **h**: high sequencing depth, high compositional variation; **h**, **l**: high sequencing depth, low compositional variation; **l**, **h**: low sequencing depth, high compositional variation; **l**, **l**: low sequencing depth, low compositional variation.

Although the results of the Jaccard index for Inv-Comp-gLASSO look quite different from the other measures at the first glance, it actually implies a similar conclusion. First, we notice that there is no Jaccard index for the first few tuning parameters that are large enough. This is because the resultant network is empty with either true or false reference. Although two empty networks agree with each other perfectly, the Jaccard index is not well defined. Then, as Inv-Comp-gLASSO starts to pick up edges when λ decreases, the Jaccard index is quite low in some settings, suggesting that the two networks are dissimilar. However, this is due to the fact the Jaccard index is a much more "strict" similarity measure than the Hamming similarity. For example, for two networks with 100 possible total edges, if both networks only have one but a different edge, then the Jaccard index is 0 while the normalized Hamming similarity is 0.98. Finally, as the networks become denser, the Jaccard index increases quickly and stabilizes at a quite high value in most settings.

It is also notable that both the Jaccard index and the normalized Hamming similarity are relatively high in the "high sequencing depth, low compositional variation" setting and relatively low in the "low sequencing depth, high compositional variation" setting, which is consistent with the finding for the normalized Manhattan similarity.

3.4.2.3 ROC Curves

An ROC curve is plotted from the average of true positive rates and the average of false positive rates over 100 replicates. An ROC curve can be regarded as an indirect measure of the invariance (two networks possessing similar ROC curves is a necessary but not sufficient condition for the two networks to be similar). However, it is crucial to evaluate the algorithms with this criterion, since it answers the question: “Does the performance of the algorithm depends on the choice of reference?”

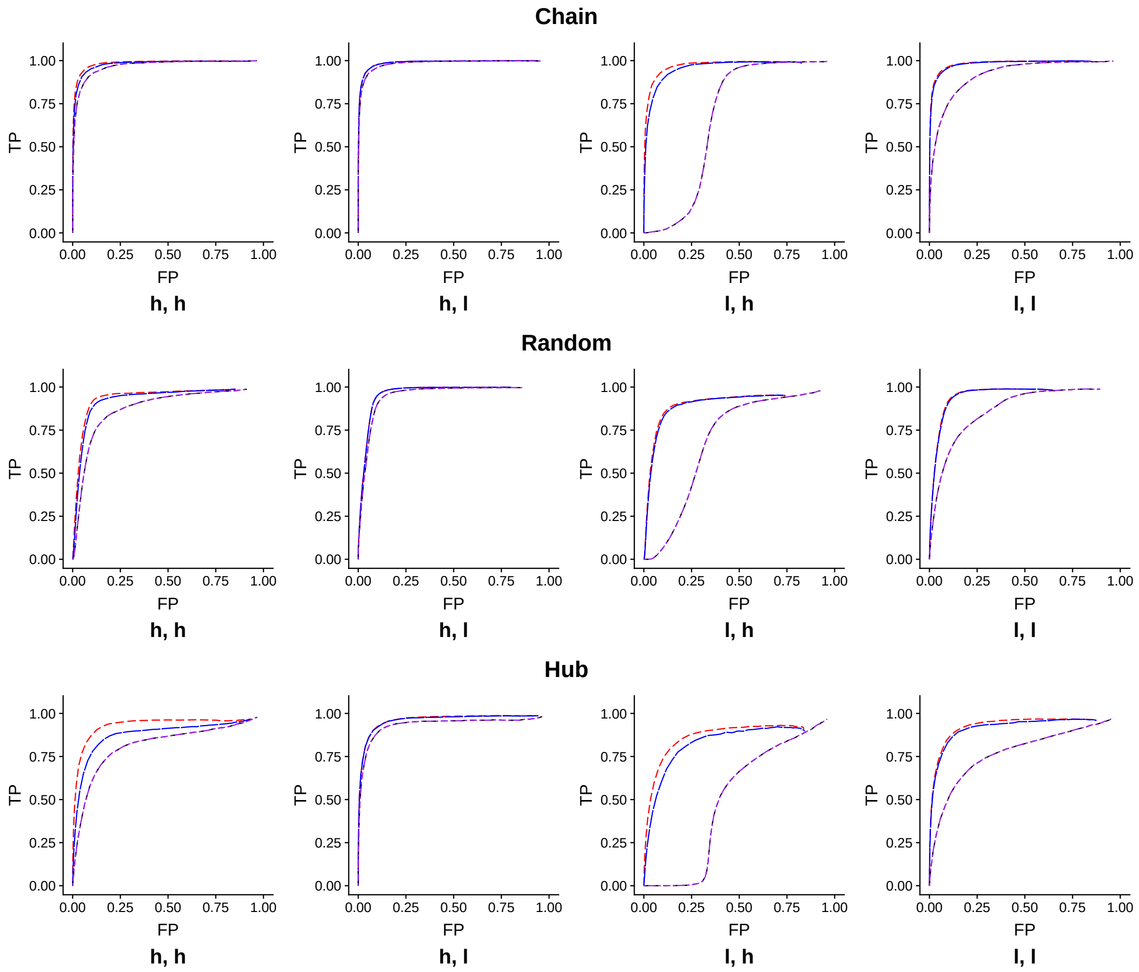

Figure 3.4: ROC curves for Inv-Comp-gLASSO and Inv-gLASSO with true and false references. Long-dashed blue: Inv-Comp-gLASSO with true reference; dashed red: Inv-Comp-gLASSO with false reference; dashed-dotted purple: Inv-Comp-gLASSO with true reference; dotted black: Inv-Comp-gLASSO with false reference. **h**, **h**: high sequencing depth and high compositional variation; **h**, **l**: high sequencing depth and low compositional variation; **l**, **h**: low sequencing depth and high compositional variation; **l**, **l**: low sequencing depth and low compositional variation.

We could see that the ROC curves from Inv-Comp-gLASSO, regardless of the choice of the reference, dominate the ones from Inv-gLASSO in all settings. The two ROC curves from Inv-gLASSO lay perfectly on top of each other, while the curves from Inv-Comp-gLASSO are also fairly close to each other. These empirical results validate the theoretical reference-invariance property for both methods. In addition, Inv-Comp-gLASSO has the most obvious advantage over Inv-gLASSO in the "low sequencing depth, high compositional variation" setting and they perform almost identically in the "low sequencing depth, high compositional variation" setting. Although the similarity measures are lower in the "most favorable" setting for Inv-Comp-gLASSO (see Sections 3.4.2.1 and 3.4.2.2), the ROC curves of the two networks from the method do not deviate too much from each other in this setting.

3.5 Real Data

To further validate the theoretical reference-invariance properties of Inv-Comp-gLASSO and Inv-gLASSO, we applied them to a dataset from the TARA Ocean project, in which the Tara Oceans consortium sampled both plankton and environmental data in 210 sites from the world oceans. The data collected was later analyzed using sequencing and imaging techniques. 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 part of the TARA Oceans project, Lima-Mendez et al. (2015) investigated the impact of both biotic and abiotic interactions in oceanic ecosystem. In this article, a literature-curated list of genus-level marine eukaryotic plankton interactions was generated by a panel of experts.

Similar to Tian et al. (2020), we focused the analysis on genus level and only kept the 81 genus involved in the literature-reported interactions. For computational simplicity, we removed the samples with too small reads (< 100). As a result, it leaves with 324 samples in the final preprocessed data. From the genera that were not reported in the literature, we chose two of them, Acrosphaera and Collosphaera, with the largest average relative abundances as the references. We then applied both Inv-Comp-gLASSO and Inv-gLASSO to the ALR-transformed data with those two references, with a common sequence of tuning parameters. For each combination of a method and a reference, we also selected a tuning parameter that corresponds to the "asymptotically sparsistent" (the sparsest estimated network in the path that contains the true network asymptotically) network by StARS (Liu et al., 2010).

Figures 3.5 presents the normalized Manhattan similarity between two estimated inverse covariance matrices with the two choices of reference. Figures 3.6 plotted the Jaccard index and the normalized Hamming similarity between the two networks represented by the two estimated inverse covariance matrices. We can see from Figure 3.5 and Figure 3.6 that, all three similarity metrics stay steadily around 1 for Inv-gLASSO throughout the sequence of λ's. This agrees with our observation in simulations where Inv-gLASSO produced numerically identical inverse covariance matrices.

For Inv-Comp-gLASSO, the similarity scores start around 1 (for normalized Manhattan similarity and normalized Hamming similarity) or non-existent (for Jaccard index) when the estimated networks are empty. Then, as λ decreases, the estimated networks become denser, and the measures start to fluctuate and decline slightly at the end. In spite

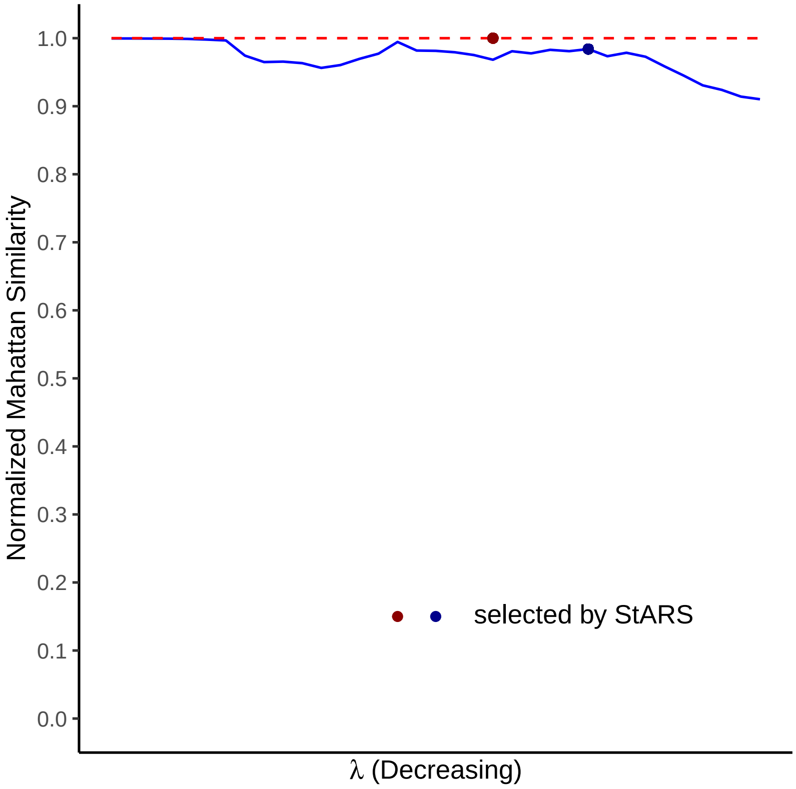

Figure 3.5: Normalized Manhattan similarity between the two estimated inverse covariance matrices with the the two choices of reference, Acrosphaera and Collosphaera, from the TARA data. Solid blue: Inv-Comp-gLASSO; dashed red: Inv-gLASSO.

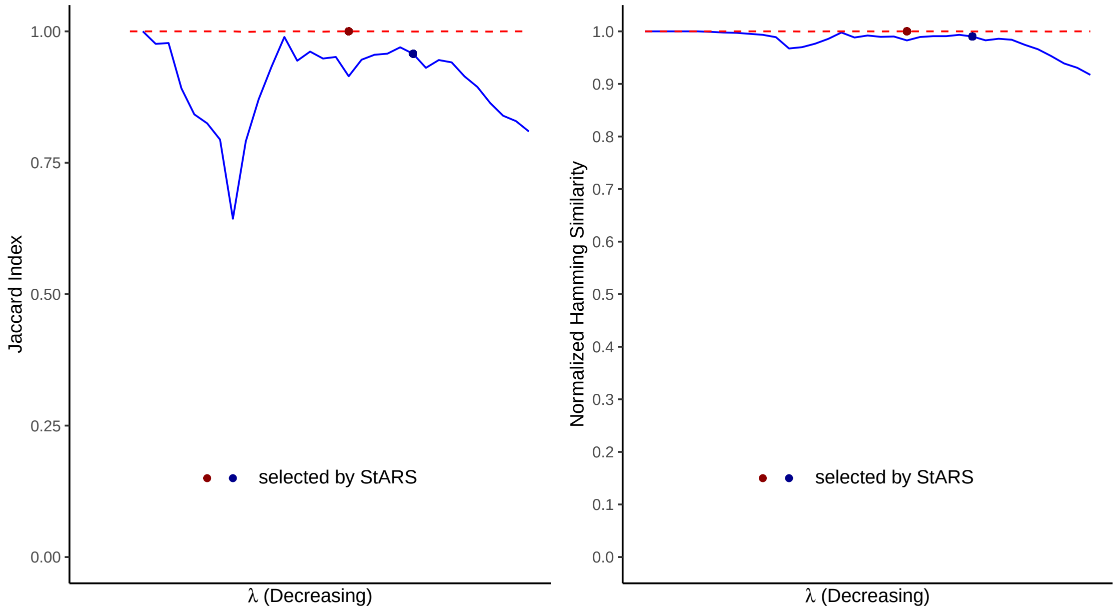

Figure 3.6: Jaccard index and normalized Hamming similarity between the two estimated networks with the the two choices of reference, Acrosphaera and Collosphaera, from the TARA data. Solid blue: Inv-Comp-gLASSO; dashed red: Inv-gLASSO.

of the fluctuations, both normalized Manhattan similarity and normalized Hamming similarity stay above 0.9, while the lowest Jaccard index is about 0.64. As discussed earlier, Jaccard index is a stricter measure than normalized Hamming similarity.

Within each method, StARS picked the same tuning parameter λ regardless of the choice of the reference, as denoted by the red dot for Inv-gLASSO and the blue dot for Inv-Comp-gLASSO in Figures 3.5 and 3.6. In other words, the red and blue dots represent the tuning parameters corresponding to the final estimated networks selected by StARS. Again, the three similarity measures for the two final inverse covariance matrices or networks from Inv-gLASSO is almost 1, while the normalized Manhattan similarity, Jaccard index, and normalized Hamming similarity are 0.98, 0.96 and 0.99 for Inv-Comp-gLASSO. All these high similarity scores imply the empirical invariance for Inv-gLASSO and Inv-Comp-gLASSO. Both methods result in invariant inverse covariance matrices and thus the corresponding networks with respect to the choices of the reference genus (Acrosphaera or Collosphaera) when applied to the TARA Ocean eukaryotic dataset.

3.6 Discussion

In this work, we established the reference-invariance property in sparse inverse covariance matrix estimation and network construction based on the ALR transformed data. Then, we proposed the reference-invariant versions of the compositional graphical lasso and graphical lasso by modifying the penalty in their respective objective functions. In addition, we validated the reference-invariance property of the proposed methods empirically by applying them to various scenarios of simulations and a real TARA Ocean eukaryotic dataset.

It is noteworthy that the reference-invariance property is a general property for estimating the inverse covariance matrix based on the ALR transformed data. We proposed reference-invariant versions of compositional graphical lasso and graphical lasso based on this property, however, one may revise other existing methods for inverse covariance matrix estimation based on the ALR transformed data. The trick is to revise the objective function so that it becomes invariant with respect to the choice of the reference. Subsequently, the resultant inverse covariance matrix and network are expected to be reference-invariant both theoretically and empirically, the latter of which may depend on the algorithm that is used to optimize the reference-invariant objective function.

4 Summary and Discussion

4.1 Summary

Our contribution is mainly twofold. We proposed a novel method called compositional graphical lasso to estimate the conditional dependence among microbes while accounting for the compositionality, finite and heterogeneous sequencing depth, and high-dimensionality of the microbiome data. Compositional graphical lasso is based on a hierarchical model called logistic normal multinomial model and the microbial interaction network is recovered by estimating a sparse inverse covariance matrix in the logistic normal distribution. Additionally, the algorithm to optimize the objective function under this model iterates between a Newton-Raphson algorithm and the graphical lasso. We established the theoretical convergence of the algorithm and showed that compositional graphical lasso outperforms its competitors under various simulation settings and in an application to the TARA Ocean eukaryotic data (Lima-Mendez et al., 2015).

Although compositional graphical lasso performs well in recovering the microbial interaction network, it remains unclear how different choices of the reference taxon would affect its resultant network. Some robustness or ideally, invariance, with respect to the choice of the reference taxon is crucial under this framework. To this end, we further established the reference-invariance property of a sub-network that corresponds to the non-candidate-reference taxa. We also proposed a reference-invariant version of compositional graphical lasso via a simple modification to the penalty in its objective function. In fact, similar modifications could be applied to other graphical model based methods such as graphical lasso. We validated the reference-invariance property of the proposed methods through simulations and an application to the same TARA Ocean eukaryotic data.

4.2 Discussion

There are a couple of topics we'd like to discuss with respect to our proposed methods and the directions that could be worth pursuing in the future.

4.2.1 Computation Time

Though we didn't evaluate the runtime of compositional graphical lasso explicitly, it is apparent that compositional graphical lasso takes more time than graphical lasso, since it incorporates the latter as an iteration step. As mentioned in Chapter 2, we also observed in simulations that it takes more iterations to converge in the settings that is more advantageous for compositional graphical lasso than the settings where compositional graphical lasso and graphical lasso share similar performance. However, in our experience, even with one iteration of compositional graphical lasso, there was already significant improvement in ROC curves compared to graphical lasso and neighborhood selection. In other words, strict convergence criterion doesn't seem to be necessary based on our experience. On the other hand, the algorithm takes more iterations to converge when the tuning parameter is small and fewer iterations to converge when the tuning parameter is reasonably large. In practice, a reasonably sparse network is of more interest corresponding to a reasonably large tuning parameter (e.g. the tuning parameter selected by StARS). Besides, the Newton-Raphson part, which is in addition to graphical lasso in compositional graphical lasso, is easily parallelizable. In summary, the computation time of compositional graphical lasso doesn't seem to be of a big concern.

4.2.2 The advantage of the ALR Transformation

One of the major differences between our work and many existing methods in the estimation of the conditional dependence microbial networks is that we used the ALR transformation instead of the CLR transformation, with the advantage of having a full-rank covariance matrix and a well-defined inverse covariance matrix in the model. Most of the previous methods get around this by imposing an L_1-norm penalty to account for the additional unknown absolute abundance parameters, while the ALR-based methods doesn't have such a burden. Major concerns regarding the ALR transformation lie in the robustness of downstream analysis with different choices of the reference taxon. We hope that our finding of the reference-invariance property has alleviated such concerns.

4.2.3 Alternative Approach

In compositional graphical lasso, we maximized the penalized joint log-likelihood function of the latent vectors $\mathbf{z}_i$'s and the inverse-covariance matrix Ω, or equivalently, the posterior distribution of $\mathbf{z}_i$'s given the observed data $\mathbf{x}_i$'s. Unfortunately, we couldn't establish the asymptotic property of the resulting estimator, due to the complexity of the parameter space induced by the the latent vectors.

A possible alternative approach could be to maximize the marginal likelihood function of the observed data $\mathbf{x}_i$'s only. Such optimization may be realized by an Monte Carlo EM (MCEM) algorithm similar to the idea described in Xia et al. (2013), which iteratively updates the expectation of Σ given the current estimate of Ω, and then calculates Ω from a graphical lasso step which takes the expectation of Σ as input. In reality, this procedure could encounter certain difficulty, as the approximation error might be dominated by the Monte Carlo (MC) error (Jiang, 2007). One possible rescue would be the automated method (Booth and Hobert, 1999) that computes the suitable MC size at each iteration of the MCEM. If the aforementioned attempt were a success, then this approach strives to empirically compute the maximum penalized likelihood estimator. One could potentially adapt the well-established asymptotic theory of maximum likelihood estimators to investigate the theoretical properties for this estimator.

Bibliography

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 139–177.

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(6):716–723.

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

Arrigo, K. R. (2004). Marine microorganisms and global nutrient cycles. *Nature*, 437(7057):349.

Ban, Y., An, L., and Jiang, H. (2015). Investigating microbial co-occurrence patterns based on metagenomic compositional data. *Bioinformatics*, 31(20):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(Mar):485–516.

Bardgett, R. D., Freeman, C., and Ostle, N. J. (2008). Microbial contributions to climate change through carbon cycle feedbacks. *The ISME journal*, 2(8):805–814.

Billheimer, D., Guttorp, P., and Fagan, W. F. (2001). Statistical interpretation of species composition. *Journal of the American statistical Association*, 96(456):1205–1214.

Booth, J. G. and Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):265–285.

Clarholm, M. (1985). Interactions of bacteria, protozoa and plants leading to mineralization of soil nitrogen. *Soil Biology and Biochemistry*, 17(2):181–187.

Cotner, J. B. and Biddanda, B. A. (2002). Small players, large role: microbial influence on biogeochemical processes in pelagic aquatic ecosystems. *Ecosystems*, 5(2):105–121.

Fang, H., Huang, C., Zhao, H., and Deng, M. (2015). Cclasso: correlation inference for compositional data through lasso. *Bioinformatics*, 31(19):3172–3180.

Fang, H., Huang, C., Zhao, H., and Deng, M. (2017). gcoda: conditional dependence network inference for compositional data. *Journal of Computational Biology*, 24(7):699–708.

Faust, K. and Raes, J. (2012). Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10(8):538.

Friedman, J. and Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS computational biology*, 8(9):e1002687.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American statistical Association*, 70(350):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(2):215–223.

Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell system technical journal*, 29(2):147–160.

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., Creasy, H. H., Earl, A. M., FitzGerald, M. G., Fulton, R. S., et al. (2012). Structure, function and diversity of the healthy human microbiome. *nature*, 486(7402):207.

Jaccard, P. (1901). Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bull Soc Vaudoise Sci Nat*, 37:241–272.

Jiang, D., Sharpton, T., and Jiang, Y. (2020). Microbial interaction network estimation via bias-corrected graphical lasso. *Statistics in Biosciences*. To appear.

Jiang, J. (2007). *Linear and generalized linear mixed models and their applications*. Springer Science & Business Media.

Kamada, N., Chen, G. Y., Inohara, N., and Núñez, G. (2013). Control of pathogens and pathobionts by the gut microbiota. *Nature immunology*, 14(7):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(10):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(5):e1004226.

Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology*, 22(1):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(6237):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*, pages 1432–1440.

Mazmanian, S. K., Round, J. L., and Kasper, D. L. (2008). A microbial symbiosis factor prevents intestinal inflammatory disease. *Nature*, 453(7195):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(3):e00031–18.

McMurdie, P. J. and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS computational biology*, 10(4).

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Meinshausen, N., Bühlmann, P., et al. (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34(3):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(302):275–309.

Nicholson, J. K., Holmes, E., Kinross, J., Burcelin, R., Gibson, G., Jia, W., and Pettersson, S. (2012). Host-gut microbiota metabolic interactions. *Science*, 336(6086):1262–1267.

Paerl, H. and Pinckney, J. (1996). A mini-review of microbial consortia: their roles in aquatic production and biogeochemical cycling. *Microbial Ecology*, 31(3):225–247.

Pearson, K. (1897). On a form of spurious correlation which may arise when indices are useed in the measurement of organs. In *Royal Soc., London, Proc.*, volume 60, pages 489–502.

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(3):1142–1149.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Smith, P., Willemsen, D., Popkes, M., Metge, F., Gandiwa, E., Reichard, M., and Valenzano, D. R. (2017). Regulation of life span by the gut microbiota in the short-lived african turquoise killifish. *elife*, 6:e27014.

Sonnenburg, J. L. and Fischbach, M. A. (2011). Community health care: therapeutic opportunities in the human microbiome. *Science translational medicine*, 3(78):78ps12–78ps12.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133.

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., et al. (2015). Structure and function of the global ocean microbiome. *Science*, 348(6237):1261359.

Tackmann, J., Rodrigues, J. F. M., and von Mering, C. (2019). Rapid inference of direct interactions in large-scale ecological networks from heterogeneous microbial sequencing data. *Cell systems*, 9(3):286–296.

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., Prill, R. J., Tripathi, A., Gibbons, S. M., Ackermann, G., et al. (2017). A communal catalogue reveals earth's multiscale microbial diversity. *Nature*, 551(7681):457–463.

Tian, C., Jiang, D., and Jiang, Y. (2020). Microbial network recovery by compositional graphical lasso. Manuscript in preparation.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494.

Wieder, W. R., Bonan, G. B., and Allison, S. D. (2013). Global soil carbon projections are improved by modelling microbial processes. *Nature Climate Change*, 3(10):909–912.

Xia, F., Chen, J., Fung, W. K., and Li, H. (2013). A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*, 69(4):1053–1063.

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

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*, 10:516.

Yuan, H., He, S., and Deng, M. (2019). Compositional data network analysis via lasso penalized d-trace loss. *Bioinformatics*, 35(18):3404–3411.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):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(Apr):1059–1062.