Inferring microbiome networks using graphical models

Jared Samilow
Brown University
Providence, RI
jared_samilow@brown.edu

Shibu Yooseph
University of Central Florida
Orlando, FL
shibu.yooseph@cs.ucf.edu

Abstract

We represent microbial count data using a Graphical Gaussian Distribution, employing the lasso technique to approximate its inverse covariance matrix. We use the Extended Bayesian Information Criterion[1] and Stability Approach to Regularization Selection to select the penalty for the lassso. To test for accuracy, we simulate data, apply our methodology to it, and compare the results to the known ground-truth.

1. Introduction

Microbes account for more than half of cells in the human body, but biologists have only just begun to catalogue them and explore their impact on health and diseases. The goal of this project is to assess the correlations between microbes – when the presence of one species predicts the presence of another. Our data come from microbiome samples from different parts of the human body. The samples were sequenced on their 16S ribosomal marker and microbiologists recorded how many times a particular species (a taxon) was observed. This leads to an $n \times p$ matrix M of n samples of p taxa which, as the title suggests, is modeled as a graph G, with the taxa the nodes of G, and the relationships between them the undirected edges.

1.1. Motivation

Microbes play critical roles in various bio-geochemical processes and elemental cycling. Microbes are also associated with other life forms, including plants and animals. Every human has several trillion microbes in contact with their body, and on-going research continues to reveal the important roles these microbes play in human health and disease (for example, see [2, 3]). By mapping their co-frequencies, we can make predictions, we can better understand microbes' effects on humans, such as, for example, when the presence of a particular taxon predicts a certain medical condition.

1.2. Formal Statement of Problem

Given a matrix of n samples and p taxa, where $M[i, j]$ denotes the relative abundance of taxon j in sample i, construct a graph, where V is the node set and $E \subset V \times V$ denotes the edge set. Node $v_i \in E$ is a random variable associated with the abundance of taxon i, and for nodes $v_i, v_j \in V$ edge (v_i, v_j) denotes the partial correlation of v_i and v_j.

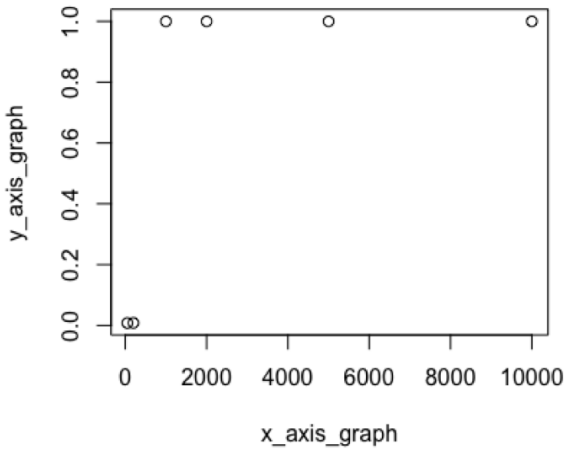


Figure 1. Plot of sensitivity scores on simulated data for different sample sizes.

1.3. Framework

Almost all programming was carried out in R. Occasionally, we turned to Matlab and Python. The R packages used were, in no particular order, `compositions`, `rlnorm`, `glasso`, `mass`, `corpcor`, `MASS` and, most significantly, `huge`.

1.4. Solution

We use a Graphical Gaussian Model (GGM) to estimate the inverse covariance matrix of G, along with the Extended Bayesian Information Criterion (EBIC) and StArs models to train the estimation techniques employed by the GGM. A crucial assumption this model makes is that the graph being estimated is sparse. This is a reasonable assumption, since we'd expect *most* microbes to exist independent of one another.

2. Methodology

2.1. Simulated Data

A challenge of inferring microbiome networks is *verification*: You cannot visually observe whether or not an inference is correct. To overcome this, we generated simulated data to serve as the ground truth, and then tested our methodology on it. We used the `huge` package to create hundreds of graphs – some random, some cluster, some hub, some band – and then estimated their adjacency and partial correlation matrices. More specifically, we

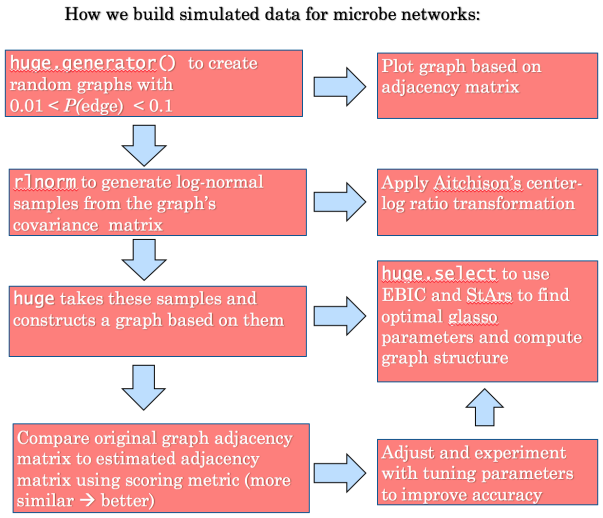


Figure 2. On the left is the "true" graph we generated. On the right is the approximation given by our methodology using EBIC. The approximation is good, but the error is intolerable and must be reduced.

1. Selected a type of graph to work with. Most of the time, we selected a random graph with $P(\text{edge}) = 0.05$ to guarantee sparsity. We also had to choose how many samples to generate from the graph, how many nodes the graph should have, etc. We generated graphs for $n = 50$ samples, 200 samples, 1,000 samples, 2,000 samples, 5,000 samples and 10,000 samples. Whichever we chose, we then obtained the covariance matrix for these samples. Generally, we made 10 graphs for each n and averaged their results to control for outliers.

2. The covariance matrix, computed by `huge.generate`, was provided to the `rlnorm` package to generate log-normal samples. We assumed distribution around a mean of 4 to increase the likelihood of obtaining positive numbers.

3. Through the `huge` package's model selection features, we used EBIC and StArs to select the penalty for the graphical lasso to estimate the inverse covariance matrix of the graph.

4. We then compared the estimated graphs to the original graphs, as dicussed below.

2.2. Real Data-Sets

We used standard and straightforward scoring techniques. To compare adjacency matrices, we simply computed the *sensitivty* and *specificity* metrics. A true positive was a correct identification of an edge, a false negative an incorrect ommission of an edge, etc. For simplicity, we did not yet consider whether a "0" was correctly identified. To evaluate the partial correlation values of the "true" graph PC_{true} against those of the estimated graph $PC_{estimated}$, we made a new matrix M where

$$M[i,j] = \ln\left(\frac{PC_{true}[i,j]}{PC_{estimated}[i,j]}\right)$$

We then simply averaged the entries of M to obtain at a numeric score. Such scores are plotted below for $n = 50, 200, 1,000, 2,000, 5,000$ and $10,000$, with $EBIC = 0.5$.

This is best explained by looking at what we did to a single data-set. As an example: We took a data-set of 39 samples of *E. andrei*, a bacterial community in the human gut. The data consisted of integer counts of different microbe species. Because such data is compositional in nature, a transformation is required to control for bias (see Aitchison, cited at end). We opted for a centered-log ratio transformation. Each entry of the matrix was "compositionalized" and divided the geometric mean of the sample it was in:

$$M[i,j] = \frac{M[i,j]}{\sum_{k \in M[i,]} M[i,k]} \cdot \left(\frac{1}{gm(M[i,])}\right)$$

where gm denotes the geometric mean. We then modeled it using the Gaussian Distribution methodology discussed above and computed the partial correlations. Results for this are discussed below.

3. Results

Tables and graphs are shown below.

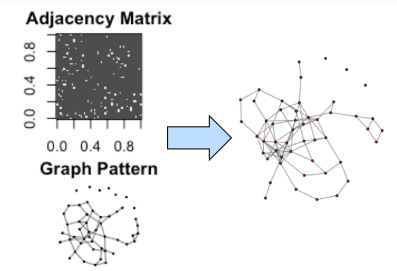


Figure 3. On the left is the ”true” graph we generated. On the right is the approximation given by our methodology using EBIC. The approximation is good, but the error is intolerable.

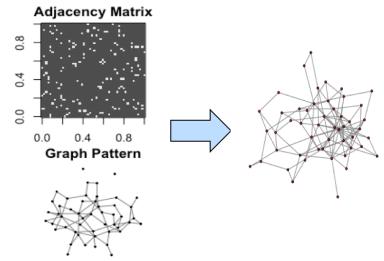


Figure 4. On the left is the ”true” graph we generated. On the right is the approximation given by our methodology using EBIC. The approximation is good, but the error is intolerable.

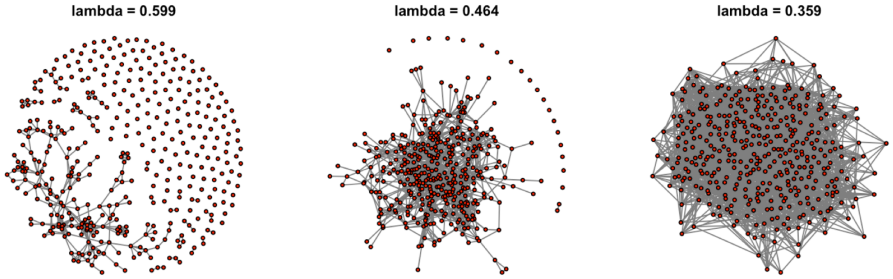


Figure 5. For n = 400 with different lambdas .

Table 1. The partial correlations of the data-set, for the 8 most frequent taxa

3.1. Simulated Data

3.2. Real Data-Sets

3.3. Discussion

The results are statistically significant and “reasonable,” i.e. they are consistent with common sense. We need to simulate and explore “hub” graphs. Additionally, we want to see how our approximation for the real-world data-set compares with the estimates given by SparCC and SPIEC-EASI, two other methods for inferring microbiome networks described by 16S ribosomal data.