Limitations of Graph Selection from High-Dimensional Compositional Data

Camden Lopez

Department of Statistics, Oregon State University, Corvallis, OR 97331, USA

email: camden.lopez@oregonstate.edu

May 2017

Abstract

Microbial abundance data produced by high-throughput genomic sequencing techniques such as 16S amplicon sequencing can be used to investigate ecological relationships among microbes. The data require special methods for analysis because they are both compositional and high-dimensional. One recently proposed method infers conditional dependence relationships within a microbial community based on log-ratio transformed abundances and graph selection via graphical lasso or neighborhood selection. The method relies on the covariance matrix for the log-ratio transformed data being approximately equal to the covariance matrix for the log-absolute abundances. Two limitations of using compositional data in this setting are that any given covariance structure for the log-ratio transformed abundances could correspond to a wide range of log-absolute abundance covariances representing diverse conditional dependence relationships, and some log-absolute abundance covariance structures can result in misleading graph inferences based on the log-ratio data.

Introduction

This work was motivated by the problem of inferring microbial interaction networks from compositional data. High-throughput sequencing methods such as 16S amplicon sequencing produce count data representing the abundances of microbial operational taxonomic units (OTUs) in a collection of samples. The counts are assumed to be proportional to the absolute abundances of the OTUs, but the sample preparation and sequencing process introduce biases such that the total sample counts do not accurately reflect the total abundance of microbes in the samples. Therefore, only relative abundances within a sample are meaningful, and the data are considered compositional (Gloor, Wu, and Pawlowsky-Glahn, 2016; Tsilimigras and Fodor, 2016). The data for each sample essentially amount to a vector of proportions which sum to one.

John Aitchison (1986) is widely cited for the theory and methodology he developed for compositional data analysis. Microbial abundance data present a new challenge in that they are high-dimensional, with the number of OTUs often greater than the number of samples, and sparse, with many zero counts.

Given a set of microbial abundance data, one goal for analysis is sometimes to determine which OTUs appear to interact or have some kind of association. Associations among variables are often quantified using correlation, but correlation analysis of proportions making up a composition can produce spurious results because of the sum-to-one constraint. Several methods, such as SparCC (Friedman and Alm, 2012), have been developed to perform correlation analysis in a manner which accounts for the compositionality.

A different approach is taken with SPIEC-EASI (Kurtz, Müller, Miraldi, Littman, Blaser, and Bonneau, 2015), which aims to infer a graph representing conditional dependence relationships among OTU absolute abundances based on relative abundance data. The proposed method involves a log-ratio transformation commonly used with compositional data. Based on an assumption that the covariance matrix of the transformed data is a good estimate of the covariance matrix of the log-absolute abundances, one of two graph selection techniques is applied to the transformed data. Here I focus on the graphical lasso technique.

Considering how relative abundances provide limited information about absolute abundances, one would expect there to be limitations in any method which claims to reliably recover absolute abundance relationships based on compositional data. I investigated the limitations of graph selection as proposed by Kurtz et al. (2015) and attempted to characterize the situations where where it works well and those where it does not.

Graph Selection from Compositional Data

Let Y be an $n \times p$ matrix of non-negative OTU count data such as that obtained from 16S amplicon sequencing, consisting of n samples and p OTUs. Element y_{ij} is the count for OTU j in sample i. The sample totals are uninformative, and one would typically normalize each sample by its total to obtain proportions $x_{ij} = y_{ij} / \sum_{j=1}^{p} y_{ij}$. Aitchison (1986) suggested that compositional data be analyzed using log-ratios, and Kurtz et al. (2015) use the centered log-ratio (clr) transformation, which has the advantage that it treats the components symmetrically (that is, it doesn't require arbitrarily choosing a reference component). For compositional vector $x = (x_1, \ldots, x_p)$ with geometric mean $g(x) = \left(\prod_{i=1}^{p} x_i\right)^{1/p}$,

$$\begin{aligned} z = \operatorname{clr}(x) &= \left(\log \frac{x_1}{g(x)}, \ldots, \log \frac{x_p}{g(x)}\right) \\ &= \left(\log x_1 - \frac{1}{p}\sum_{i=1}^{p} \log x_i, \ldots, \log x_p - \frac{1}{p}\sum_{i=1}^{p} \log x_i\right) \end{aligned} \tag{1}$$

The result of the clr transformation is invariant to arbitrary scaling of the input vector, so that if x is a vector of relative proportions obtained from observed abundances y, which are proportional to absolute abundances w, then $\operatorname{clr}(x) = \operatorname{clr}(y) = \operatorname{clr}(w)$.

Let Z be the matrix obtained by clr transformation of the rows of Y. (Sequencing data typically have many zero counts which would make the clr transformation impossible, so it is necessary to add a pseudo count of 1 or by some other means make all abundances strictly positive.) A row of Z represents the distances of each OTU's abundance in a sample from the mean of the sample, on the log scale, and it contains all of the information that the original abundance observations contained, assuming the total observed abundance was uninformative.

A microbial interaction network can be inferred based on the graphical model representation of the log-absolute microbial abundances. Let $W = (W_1, \ldots, W_p)$ be a random vector of strictly positive absolute abundances. The undirected graphical model representation of $\log W = (\log W_1, \ldots, \log W_p)$ has one node for each $\log W_i$, $i = 1, \ldots, p$, and an edge between node i and node j, $i \neq j$, if and only if $\log W_i$ and $\log W_j$ are dependent (the distribution of one depends on the value of the other) conditional on the values of $\{\log W_k : k \neq i, k \neq j\}$. The lack of an edge between nodes i and j indicates conditional independence between $\log W_i$ and $\log W_j$.

Following the notation of Aitchison (1986), let $\Omega = \mathrm{cov}(\log W)$ be the population covariance matrix for the log-absolute abundances and $\Gamma = \mathrm{cov}(Z) = \mathrm{cov}(\mathrm{clr}\, W)$ be the covariance matrix for clr-transformed abundances. If $\log W \sim \mathrm{N}(\mu, \Omega)$, then $\log W_i$ and $\log W_j$ are conditionally dependent if and only if the (i, j) entry of Ω^{-1} is non-zero, with the sign of the entry corresponding to a linear conditional relationship of the opposite sign. The graphical lasso (Friedman, Hastie, and Tibshirani, 2008) is one method for selecting non-zero entries of Ω^{-1} in the high-dimensional setting (possibly $n < p$) when Ω^{-1} is assumed to be sparse.

Graphical lasso estimates a sparse Ω^{-1} by minimizing the multivariate normal log-likelihood with a penalty on the l_1 norm of Ω^{-1}. The graphical lasso estimate is

$$\widehat{\Omega^{-1}}_{\mathrm{glasso}} = \operatorname*{arg\,min}_{\Omega^{-1}} \Big[-\log\det(\Omega^{-1}) + \mathrm{tr}(\Omega^{-1}\hat{\Omega}) + \lambda \|\Omega^{-1}\|_1 \Big] \tag{2}$$

with Ω^{-1} restricted to positive definite matrices. $\hat{\Omega}$ is the maximum likelihood estimate of Ω, $\|\Omega^{-1}\|_1 = \sum_{i,j} |(\Omega^{-1})_{ij}|$ is the l_1 norm of Ω^{-1}, and λ is the so-called lasso penalty parameter controlling the sparsity of $\widehat{\Omega^{-1}}_{\mathrm{glasso}}$.

With only the compositional data Y and its clr form Z at hand, $\hat{\Omega}$ is unavailable. However, for many settings, $\hat{\Omega} \approx \hat{\Gamma}$, where $\hat{\Gamma}$ is the estimate of Γ calculated from Z in the same way that $\hat{\Omega}$ would be calculated from the log-absolute abundances, were they available. Substituting $\hat{\Gamma}$ for $\hat{\Omega}$ into (2) results in the sparse estimate of Ω^{-1} that will be considered here:

$$\widehat{\Omega^{-1}} = \operatorname*{arg\,min}_{\Omega^{-1}} \Big[-\log\det(\Omega^{-1}) + \mathrm{tr}(\Omega^{-1}\hat{\Gamma}) + \lambda \|\Omega^{-1}\|_1 \Big] \tag{3}$$

Conditional dependence relationships among the log-absolute abundances are inferred based on the non-zero entries of $\widehat{\Omega^{-1}}$.

Covariance Properties

From (1), $\mathrm{clr}(W) = \mathrm{G} \log W$ where $\mathrm{G} = \mathrm{I}_p - \frac{1}{p}\mathrm{J}$, I_p is the $p \times p$ identity matrix, and J_p is the $p \times p$ matrix of all ones. Therefore,

$$\Gamma = \mathrm{cov}(\mathrm{G} \log W, \mathrm{G} \log W) = \mathrm{G}\Omega\mathrm{G} \tag{4}$$

The same relation holds for empirical covariances: $\hat{\Gamma} = \mathrm{G}\hat{\Omega}\mathrm{G}$. In the following discussion, Ω and Γ can be understood to refer to either empirical or population covariances. The entries of Γ are related to the entries of Ω by

$$\gamma_{ij} = \omega_{ij} - \overline{\omega}_{i\cdot} - \overline{\omega}_{j\cdot} + \overline{\omega}_{\cdot\cdot} \tag{5}$$

where $\overline{\omega}_{i\cdot}$ is the average of entries in the ith row of Ω, and $\overline{\omega}_{\cdot\cdot}$ is the average of all entries of Ω. Equation (5) makes it evident that all rows and columns of Γ sum to zero. Ω has $p(p+1)/2$ free parameters. As a result of the sum-to-zero constraint, Γ has p fewer parameters, for a total of $p(p-1)/2$.

As noted above, the estimate $\widehat{\Omega^{-1}}$ relies on the approximation $\hat{\Omega} \approx \hat{\Gamma}$. Some simple examples illustrate how this approximation depends on the structure of Ω. When the log abundances are all uncorrelated and the marginal variances are all equal to ω, $\overline{\omega}_{i\cdot} = \overline{\omega}_{j\cdot} = \overline{\omega}_{\cdot\cdot} = \frac{1}{p}\omega$ and $\gamma_{ij} - \omega_{ij} = -\frac{1}{p}\omega$ for all i and j. The correlation among the log abundances induced by the clr transformation is only $-\frac{1}{p}$, so for large p, the effect of clr transformation is negligible.

Suppose one variance is larger than the others. Let $\omega_{11} = M\omega$, with $M > 1$. When the variance in one abundance variable is so much larger than the others that it dominates the variance in the total abundance, there will be a negative correlation induced between that variable and the others when considered as a composition or following a clr transformation. For $p = 64$, the induced correlation between the variance-dominating component and the others (that is, the correlation obtained from converting Γ to the corresponding correlation matrix) is -0.08 when $M = 25$, -0.16 when $M = 100$, and -0.30 when $M = 400$.

Under some assumptions about how many strong correlations there are among the log abundances, the approximation $\Omega \approx \Gamma$ improves as p increases. If the sums of the rows of Ω grow slower than p as p increases, then each $\overline{\omega}_{i\cdot}$ and $\overline{\omega}_{\cdot\cdot}$ diminish so that $|\gamma_{ij} - \omega_{ij}|$ also diminishes. In the microbial setting, it may be reasonable to assume that for any particular OTU, the variance of its log abundance plus the covariances between it and other OTUs grows slower than p as more OTUs are included in the data.

If all $\overline{\omega}_{i\cdot}$ are approximately equal to $\overline{\omega}_{1\cdot}$, then $\gamma_{ij} - \omega_{ij} \approx -\overline{\omega}_{1\cdot}$, and if $\overline{\omega}_{1\cdot}$ is positive and non-negligible compared to the variances $\omega_{11}, \ldots, \omega_{pp}$, then this can represent a substantial distortion of the covariances. In simulations described later, it was found that graph selection performs worse in this type of situation.

Another situation where the covariances become substantially distorted is when some $\overline{\omega}_{i\cdot}$ is much larger than the others. This occurs when, for example, one variance ω_{ii} is much larger than the others, so that the total abundance in each realization of $\log W$ is strongly influenced by the value of $\log W_i$. In that case,

$\gamma_{ij} - \omega_{ij}$ tends to be a relatively large negative value for all $j = 1, \dots, p$. If $\log W_i$ and $\log W_j$, $i \neq j$, are uncorrelated, then the clr transformation induces a negative correlation in $\operatorname{clr} W_i$ and $\operatorname{clr} W_j$. This negative correlation would be observed in the proportions X_i and X_j as well (and is the type of spurious correlation which methods such as SparCC are designed to circumvent). When the abundances become compositional instead of absolute, if one component dominated the basis variance, then its abundance relative to other components will also dominate the composition and induce negative correlations.

Alternative Basis Covariances

Let $\mathcal{S}_p$ be the set of symmetric $p \times p$ matrices with real entries, which is a space with $p(p+1)/2$ dimensions. Let $\mathrm{T}_p : \mathcal{S}_p \to \mathcal{S}_p$ be the transformation defined by $\mathrm{T}_p(\Omega) = \mathrm{G}\Omega\mathrm{G}$. T_p is a linear transformation with rank $p(p-1)/2$. For each Γ representing the covariance matrix for clr-transformed abundances, there is not a unique Ω for which $\Gamma = \mathrm{T}_p(\Omega)$ but rather a p-dimensional space of such Ω matrices. We can characterize that space with the kernel of T_p, $\ker(\mathrm{T}_p) = \{\Psi \in \mathcal{S}_p : \mathrm{G}\Psi\mathrm{G} = 0_{p\times p}\}$, which, it turns out, consists of matrices with the property that the (i, j) entry is equal to the average of the ith diagonal and the jth diagonal entries.

$$\ker(\mathrm{T}_p) = \left\{ \Psi \in \mathcal{S}_p : \psi_{ij} = \frac{1}{2}(\psi_{ii} + \psi_{jj}) \right\} \tag{6}$$

For any $\Omega \in \mathcal{S}_p$ and $\Psi \in \ker(\mathrm{T}_p)$, $\mathrm{T}_p(\Omega + \Psi) = \mathrm{T}_p(\Omega) + 0^{p\times p} = \Gamma$. So for a given Ω or $\Gamma = \mathrm{T}_p(\Omega)$, the set of all matrices which T_p sends to the same Γ is characterized by adding to Ω or Γ any $p \times p$ matrix with entries equal to the average of the corresponding diagonal entries.

The only assumption about the matrices so far has been that they are symmetric. Since only valid covariance matrices are of interest, the space of matrices sent to the same Γ must be restricted to positive definite or positive semi-definite matrices.

With a compositional data set in hand, $\hat{\Gamma}$ can be calculated and used to infer a graph via graphical lasso (or neighborhood selection). The limitation of having compositional data rather than absolute abundance data shows up in the fact that $\hat{\Gamma}$ could correspond to infinitely many covariance matrices for the log-absolute abundance data, were it available. Equation (6) provides a way to explore those potential log-absolute abundance covariances, provided that one makes sure each potential $\hat{\Omega}$ is positive (semi-)definite.

For a given Γ or Ω, let Σ be in the set of matrices such that $\mathrm{T}_p(\Sigma) = \mathrm{T}_p(\Omega) = \Gamma$. It will be convenient to be able to choose some hypothetical log-abundance variances $\sigma_{11}, \ldots, \sigma_{pp}$ and calculate the remaining entries of Σ required so that $\mathrm{T}_p(\Sigma) = \Gamma$. Using the fact that $\Sigma = \Gamma + \Psi$ for some $\Psi \in \ker(\mathrm{T}_p)$, we have $\Psi = \Sigma - \Gamma$, which implies that $\psi_{ii} = \sigma_{ii} - \gamma_{ii}$ and $\psi_{ij} = \frac{1}{2}\left[(\sigma_{ii} - \gamma_{ii}) + (\sigma_{jj} - \gamma_{jj})\right]$. Then

$$\sigma_{ij} = \gamma_{ij} + \frac{1}{2}(\sigma_{ii} - \gamma_{ii} + \sigma_{jj} - \gamma_{jj}) \tag{7}$$

To put this in matrix form, let $d = (\sigma_{11} - \gamma_{11}, \ldots, \sigma_{pp} - \gamma_{pp})$ and let R be the matrix with p rows identical to d and C be the matrix with p columns identical to d^{T}. Then

$$\Sigma = \Gamma + \frac{1}{2}(\mathrm{R} + \mathrm{C}) \tag{8}$$

Now I consider where Σ is positive definite and how the log-abundance relationships could vary among the different Σs associated with a particular Γ.

Simplest Case: $p = 2$

For the case of $p = 2$ it is easy to characterize and visualize the region where Σ matrices associated with a particular Γ are positive definite. It is also easy to determine how the log-abundance relationship varies within that region.

For $p = 2$, Γ is fully determined by $\gamma_{11} = \gamma_{22} = -\gamma_{12}$, and

$$\gamma_{11} = \frac{1}{4}(\omega_{11} - 2\omega_{12} + \omega_{22}) = \frac{1}{4}\,\mathrm{var}(\log X_1 - \log X_2) = \frac{1}{4}\,\mathrm{var}\left(\log \frac{X_1}{X_2}\right) \tag{9}$$

So γ_{11} characterizes how much variation there is in the relative proportions X_1 and X_2. Now suppose we want to find Σ with diagonal entries σ_{11} and σ_{22}. Equation (7) gives

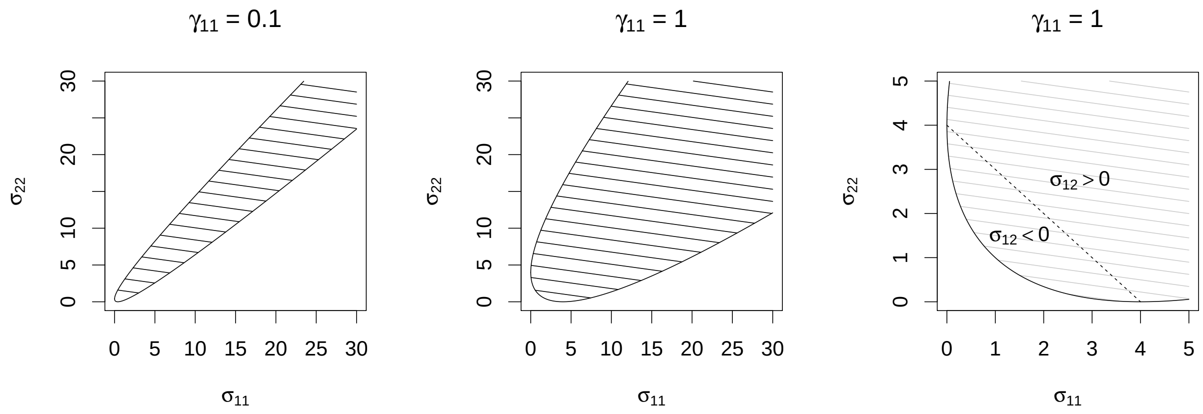
$$\sigma_{12} = \frac{1}{2}(\sigma_{11} + \sigma_{22}) - 2\gamma_{11} \tag{10}$$

When Σ is 2×2, Σ is positive definite if and only if $\det(\Sigma) > 0$. This results in the following inequality which defines the region of positive definite Σ matrices, given γ_{11}, in terms of σ_{11} and σ_{22}:

$$\left(\sqrt{\sigma_{11}} - 2\sqrt{\gamma_{11}}\right)^2 < \sigma_{22} < \left(\sqrt{\sigma_{11}} + 2\sqrt{\gamma_{11}}\right)^2 \tag{11}$$

The intuitive reason for there being bounds for σ_{22}, given σ_{11}, and vice-versa, is that given γ_{11}, we know that the differences $\log X_1 - \log X_2$ have a certain amount of variation. Suppose we hypothesize a large

Figure 1: Region of positive definite Σ for $p = 2$.



Left and center: Region (shaded) where Σ is positive definite, for $\gamma_{11} = 0.1$ and $\gamma_{11} = 1$. Right: Subregions where $\log X_1$ and $\log X_2$ are negatively correlated, uncorrelated, and positively correlated. Along the dashed line, $\sigma_{12} = 0$.

value for $\sigma_{11} = \text{var}(\log X_1)$. We can imagine setting the geometric mean $g(X) = \frac{1}{2}(\log X_1 + \log X_2)$ of each sample to achieve that variation in $\log X_1$. But the $\log X_2$ values are "tied" to the $\log X_1$ values to a degree characterized by γ_{11}. A large variance in $\log X_1$ will require that there is large variance in $\log X_2$ also. If we hypothesize σ_{11} to be smaller, then σ_{22} can be smaller also.

The region where Σ is positive definite is plotted in Figure 1 for different values of γ_{11}. We can ask of this region of alternative covariance matrices, where in the region is the relationship between $\log X_1$ and $\log X_2$ positive or negative? Where are $\log X_1$ and $\log X_2$ independent? In the $p = 2$ case, the conditional and unconditional dependence relationships between $\log X_1$ and $\log X_2$ are the same, so we can simply examine σ_{12}. From equation (10), we find that the boundary between the regions where $\sigma_{12} < 0$ and $\sigma_{12} > 0$ is the line

$$\sigma_{22} = 4\gamma_{11} - \sigma_{11} \tag{12}$$

We can compare that line to the "vertex" of the region, $(\gamma_{11}, \gamma_{11})$, which is found by solving for the point where the lower boundary intersects $\sigma_{22} = \sigma_{11}$. The vertex corresponds to $\Sigma = \Gamma$. As shown in Figure 1, for any γ_{11}, there are always regions where $\sigma_{12} < 0$, $\sigma_{12} = 0$, and $\sigma_{12} > 0$.

Near or at the boundary of the region, the correlation is near or equal to ±1. Along the line $\sigma_{22} = \sigma_{11}$, the correlation goes from -1 at $(\gamma_{11}, \gamma_{11})$ to 0 at $(2\gamma_{11}, 2\gamma_{11})$ and approaches $+1$ as $\sigma_{11} = \sigma_{22}$ increases.

Generalizing to Larger p

Characterizing the region where Σ is positive definite is more difficult with $p > 2$, but the region can be explored for a particular Γ by trial-and-error, choosing values for $\sigma_{11}, \ldots, \sigma_{pp}$ and checking whether the minimum eigenvalue of Σ is positive (indicating positive definiteness).

It appears that Σ is positive definite for $(\sigma_{11}, \ldots, \sigma_{pp}) = (\gamma_{11} + \delta, \ldots, \gamma_{pp} + \delta)$ with any $\delta > 0$. For very small δ, the entries of Σ^{-1} are all or nearly all negative, corresponding to positive conditional relationships. As δ increases, the entries of Σ^{-1} become all or nearly all positive, corresponding to negative conditional relationships. This pattern is similar to the pattern found in the $n = 2$ case.

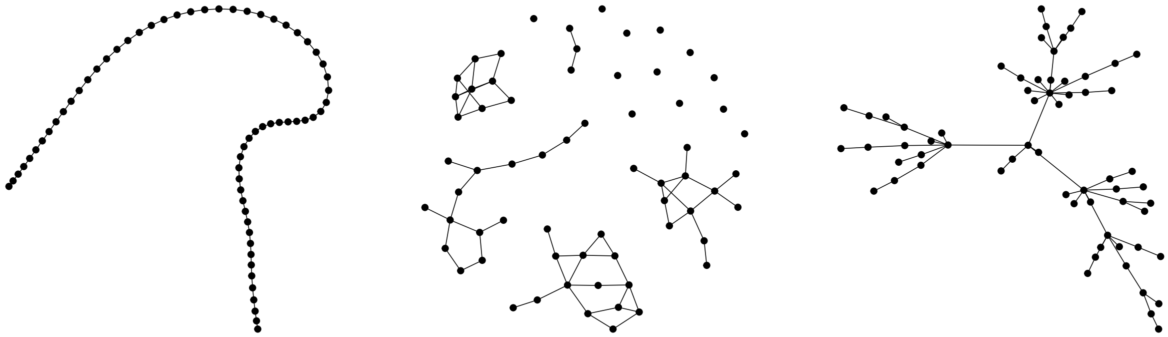
It also appears to be true for larger p that for fixed values of all σ_{ii} except one, say σ_{jj}, Σ is positive definite only for a fairly small interval of σ_{jj} values.

What can be learned from these investigations is that for any given Γ (or $\hat{\Gamma}$), there are potential log-absolute abundance covariance matrices which exhibit a wide variety of conditional and unconditional relationships among the variables and all could have resulted in Γ. What one would hope is that any valid Σs which are quite different from Γ represent implausible covariance structures, and that the Σs which are plausible are also very similar to Γ so that when graph selection is applied to Γ, its inferences are similar to what one would obtain if the absolute abundances were available. However, one may have little idea about which covariance structures are plausible or not.

Graph Selection Performance

To assess the performance of graph selection under ideal conditions, I ran simulations using log-normal absolute abundances so that the log-absolute abundances had a multivariate normal distribution. For each of $p = 64$ and $p = 256$, samples were generated using a fixed precision matrix with the locations of non-zero entries determined by one of three graph structures: band, cluster, and scale-free. The graphs for $p = 64$ are shown in Figure 2. To control the strengths of the conditional relationships, all diagonal entries in the precision matrix were 4 and all off-diagonal entries were ±1. Thus, the partial correlations were all equal

Figure 2: Band, cluster, and scale-free graph structures used in $p = 64$ simulations.



to ±0.25. The signs of the precision matrix entries were either all negative, half negative, or all positive. As noted above, a negative entry in the (i, j) position of the precision matrix corresponds to a positive conditional dependence relationship between variables i and j. That is, for fixed values of all other variables, variables i and j will have positive correlation when the entry in the precision matrix is negative and negative correlation when the entry is positive.

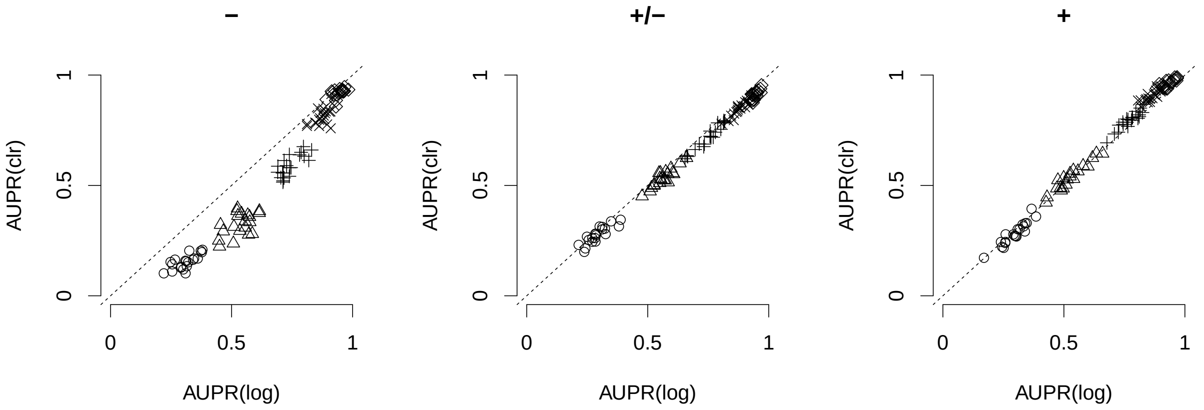
Performance was measured by the area under the precision-recall curve along the sequence of graphical lasso solutions for a decreasing sequence of λ values. Precision and recall are calculated based on the numbers of true positives (TP), false positives (FP), and false negatives (FN) as expressed below.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{13}$$

Here, a true positive is when an edge exists in the true graph and is selected by graphical lasso. A false positive is when the edge does not exist but is selected, and a false negative is when the edge does exist but is not selected. For comparison, AUPR was also measured for graphical lasso performance using the log-abundance data before transformation.

This setup for the simulations was based on the simulations reported by Kurtz et al. (2015). The differences are that their simulations used count data mimicking microbial OTU counts from a real data set, they controlled the strength of OTU relationships using the condition number of the covariance matrix, and for every precision matrix entry in all trials, the sign was made positive with probability $\frac{1}{2}$ and negative

Figure 3: Graph selection performance with $p = 64$ and a scale-free graph structure.



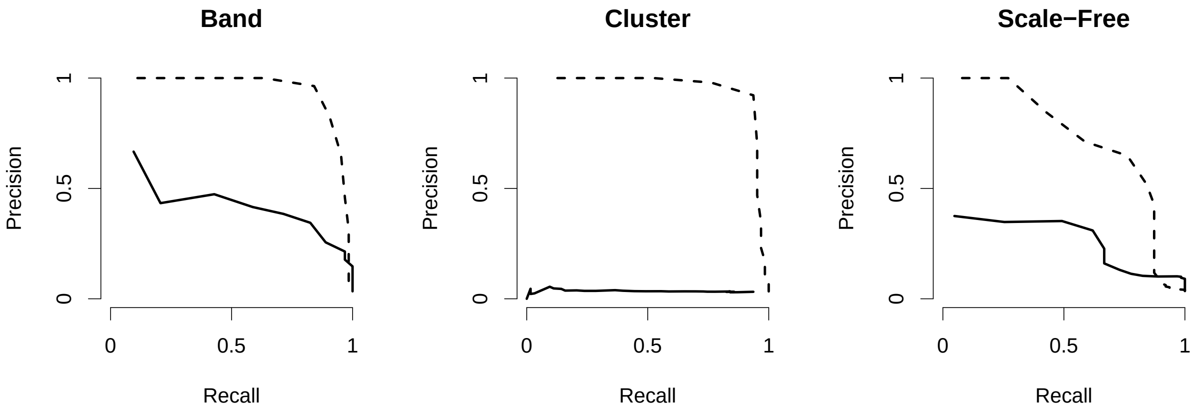
otherwise. I observed that using their graph and precision matrix generator, for a fixed condition number, the partial correlations among variables tended to be much smaller for scale-free graphs compared to band and cluster graphs. I believe this explains the difference in performance among the graph types that they reported.

Besides sample size (performance steadily improved for both the log and clr data as n increased from 32 to 512), the signs of the precision matrix entries had the largest effect on the performance of the graph selection. For $p = 64$, when all of the precision matrix entries were negative, graph selection using the clr abundance data underperformed compared to graph selection using the log abundances. The scale-free AUPR comparison is shown in Figure 3. However, for $p = 256$, performance is at least as good with clr data as with log data. Surprisingly, performance is sometimes slightly better with clr data for large n.

When all of the off-diagonal entries of the precision matrix are negative, all of the conditional relationships are positive, and the covariance matrix tends to have more and larger positive entries. As noted above, this results in a larger difference between Γ and Ω, which seems to result in worse graph selection performance, though the effect diminishes with larger p.

The conditions in these simulations were ideal in that Γ does not differ from Ω in a way that affects graph selection too strongly. Essentially, the clr data are not too different from the log data. I performed additional simulations to investigate cases where the compositional effect is stronger. For the same graphs

Figure 4: Graph selection performance in the presence of one variance-dominating variable.



Precision vs. recall for graph selection from the clr abundance data (—) *and log abundance data* (- - -).

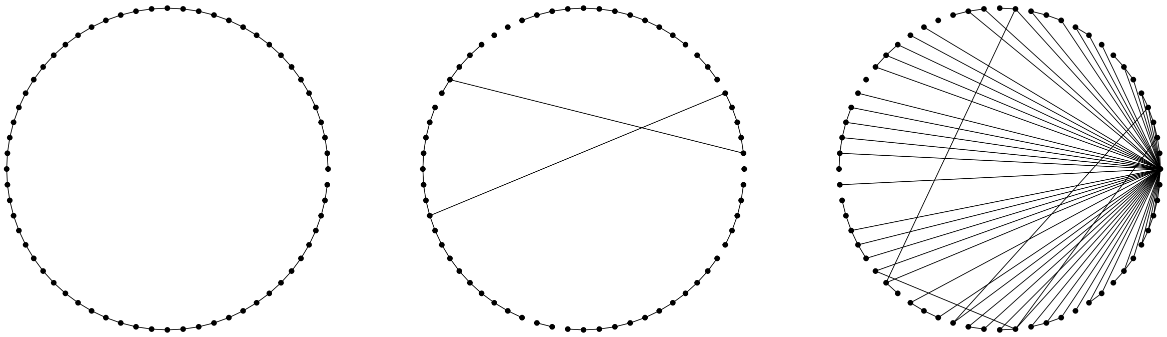
used in the $p = 64$ simulations above, I set the variance for $\log W_1$ to be 400 times as large as the other variances.

Under these conditions, where the variance of $\sum_{i=1}^{p} \log W_i$ is dominated by $\log W_1$, graph selection performs much worse on the clr data than on the log data. Graphical lasso selects many spurious edges between $\log W_1$ and the other $\log W_i$ because the clr transformation induces substantial changes in the covariances. Precision-recall curves for each graph structure and $n = 256$ are shown in Figure 4. The edges selected in the case of the band graph using the log data are compared to the edges selected using the clr data in Figure 5.

Discussion

Compositional data derived from a basis of absolute abundances are fundamentally limited in the information they provide about the absolute abundances. Any analysis which is based on covariances in the compositional data, as is the case with the graph selection method considered here, must be used with the recognition that covariances in the compositional data may differ substantially from covariances in the basis abundances. I have shown that a variety of basis abundance structures exhibiting very different types of relationships among the variables can result in the same covariance structure in the clr-transformed compositional data.

Figure 5: Spurious edges selected by graphical lasso.



Left: The true graph structure. Center: Edges selected by graphical lasso from the log abundance data, with lasso parameter $\lambda = 0.212$. Right: Edges selected from the clr abundance data with the same lasso parameter.

Based on simulations, graph selection using clr-transformed compositional data works as well as (and occasionally even better than) the graph selection that would result from the log abundances for large p, provided that the compositional effect is not too strong. Strong compositional effects result when, for example, one variable or a group of variables dominates the variance of the sample totals. It is information about the sample totals which is lost when absolute abundances become compositional abundances. Therefore, any basis covariance patterns where the sample total is substantially dependent on the abundance patterns of a subset of variables can result in poor graph selection performance. If dependence relationships among variables are fairly sparse, and if the variances of the variables do not differ much, than moderate-to-large p is sufficient for the compositional effects to be diminished or negligible.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Chapman and Hall, London, UK.

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

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

Gloor, G. B., Wu, J. R., and Pawlowsky-Glahn, V. (2016). It's all relative: analyzing microbiome data as compositions. *Annals of Epidemiology* **26,** 322–329.

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Computational Biology* **11,** e1004226.

Tsilimigras, M. C. and Fodor, A. A. (2016). Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Annals of Epidemiology* **26,** 330–335.