

Analyzing Data with Graphs: Metagenomic Data and the Phylogenetic Tree*

Elizabeth Purdom
U.C., Berkeley, Division of Biostatistics
Department of Statistics
University of California at Berkeley
367 Evans Hall #3860
Berkeley, CA 94720-3860
E-mail: epurdom@stat.berkeley.edu

October 15, 2008

Abstract

In biological experiments, researchers often have information in the form of a graph that supplements observed numerical data. Incorporating the knowledge contained in these graphs into an analysis of the numerical data is an important and non trivial task. We look at the example of metagenomic data – data from a genomic survey of the abundance of different species of bacteria in a sample. Here, the graph of interest is a phylogenetic tree depicting the interspecies relationships among the bacteria species. We demonstrate that analysis of the data in a non-standard inner-product space effectively uses this additional graphical information and produces more meaningful results.

1 Introduction

Relationships amongst either observations or variables are often conveniently summarized by a graph. Incorporating this outside information into the analysis of numerical data is of increasing interest, particularly in biology where many known properties of genes and proteins are described by complicated graphs. A common situation is to have numerical data from an experiment which is of primary interest and also additional knowledge in the form of a graph relating our observations or variables from the experiment. We would like to incorporate the information in the graph with our analysis of the experimental data. One example is data on a set of genes from a microarray experiment where there is also an established network of relationships among the genes. By including the graphical information directly in our analysis, we constrain the space of possible solutions to those that are relevant from the point of view of the known information.

The specific type of graph which we consider here is a phylogenetic tree. A phylogenetic tree is a ubiquitous graph in biology that describes the evolutionary relationship between a set of species. We

*The author would like to thank Susan Holmes for many helpful conversations and reviews of previous drafts, Elisabeth Bik, Les Dethlefsen, Paul Eckburg, and David Relman for discussions regarding the data and data analysis and for use of the data.

are motivated to consider this graph by our work with [Eckburg et al. \(2005\)](#) analyzing differences in bacterial composition based on a genomic inventory of different samples. Such “metagenomic” studies are a popular technique for measuring bacterial content. As we argue below, using phylogenetic information regarding the discovered bacteria is key in creating a meaningful analysis – particularly because of the small sample size relative to the number of bacteria found.

There are numerous different strategies for using graphical information, such as Bayesian networks and differential equation modeling; they require varying degrees of specificity in the graphical information. We focus here on a technique that is simple to implement and uses the graph to define a non-standard inner-product space in \mathbb{R}^p to perform the analysis of the numerical data.

The original analysis in [Eckburg et al. \(2005\)](#) used the technique of Double Principal Components Analysis (DPCoA) ([Pavoine et al., 2004](#)) to analyze the bacterial communities; DPCoA is a form of dimensionality reduction that incorporates *a priori* dissimilarity information about one set of factors in a contingency table. We recast this technique as an analysis based on a non-standard inner-product space defined by the outside similarity information – the phylogenetic information in the case of the bacterial analysis. When viewed in this light, the technique now has general application, since in many situations, heterogenous information can be similarly introduced into an analysis in this way.

The layout of the paper is as follows. First we will introduce the motivating example of bacterial composition in more detail before we turn to the analysis performed in [Eckburg et al. \(2005\)](#) for the bacterial data. We first review how PCA can be succinctly reformulated for non-standard inner-products. We then turn our attention to the DPCoA analysis of the bacterial composition data in [Eckburg et al. \(2005\)](#) and relate it to the PCA framework. In doing so, we recast DPCoA into a general framework that allows for ease of interpretation and comparison. The rest of the paper delves further into the implications of incorporating outside graphical information through the use of such a metric space. In particular, we propose an appropriate metric for a phylogenetic tree and evaluate the implications of that choice in the final data analysis. Throughout, we focus on the example of the phylogenetic tree and metagenomic data to illustrate the concepts. The same basic approach can be useful in including non-standard forms of knowledge – other types of graphical information in particular.

2 Motivating Example

Many types of bacteria naturally live in the bodies of humans, in particular in the intestinal tract where they can serve important functions in digestion. In [Eckburg et al. \(2005\)](#) the broad goal was to describe the kinds of bacteria found in the intestinal tract and compare the bacterial communities found in different people. To that end, each of the three patients in the study had biopsies taken at six locations in his/her colon in addition to providing a stool sample. Each of these seven samples (per patient) was then subjected to genomic techniques to try to quantify the different types of bacteria as well as their level of abundance.

However, counting and differentiating bacteria is not a simple task. The very notion of “species” (often defined as the ability to interbreed) may be vague for bacteria. Traditional techniques for identifying bacterial species grow the bacteria in a culture and then classify the bacteria as a species based on any observable characteristics as well as the nutrients needed for it to grow. This gives only limited ability to assess the present of different types of bacteria.

With the increased ease of DNA sequencing, researchers interested in bacterial communities now use “metagenomics” techniques to study all bacteria in a location based on their DNA ([Committee on Metagenomics, 2007](#)). Such genomic characterization of the present bacteria results in a measure of similarity be-

tween bacteria which is used as a surrogate for species identification. However, the use of similarity as a proxy for species creates an additional level of approximation in our analysis. The phylogenetic relationship between the species gives important outside information that will assist us in countering this problem.

2.1 Genomic techniques used to Quantify Bacteria Abundance

Genomic methods to isolate and count the bacteria do not rely on sequencing the entire genome of all the bacteria, but just a particular gene that is found in all bacteria, specifically one that encodes for a small-subunit ribosomal RNA gene (16S rDNA). Since each bacteria cell will have the same number of copies of this gene, the basic idea is to isolate from all the bacteria the DNA strands corresponding to the gene, to count different versions of the sequence, and then to identify to which bacteria the versions correspond; in this way the types and abundance of different bacteria cells in a sample are determined. Of course, since individual DNA from individual cells cannot be separately isolated, these steps must be done in mass. The main steps are

1. Isolate from each sample all bacterial copies of DNA encoding for the gene.
2. Randomly sample from the pooled copies of the gene's DNA and sequence them to completion.
3. Based on the sequence similarity of the gene, classify each copy as representing the same "species". For example, the rule in [Eckburg et al. \(2005\)](#) for grouping sequences into one "species" required all pairs in the group to have a minimum of 99% sequence similarity.
4. Count the number of times each "species" occurs in each sample.

(we refer the reader to [Eckburg et al. \(2005\)](#) for more details of the experimental process). Note that the cutoff of 99% is not a biological determinant: for different types of bacteria, different levels of sequence variability might be found amongst different species. For this reason we use the term "phylotype" rather than "species" for the classifications made by sequence similarity.

The study of [Eckburg et al.](#) of the colon of different patients was an exploratory study. It was the first effort to genomically identify the bacterial composition of the colon that compared between individuals and/or locations of the sample (many genomic experiments of this type either sampled only one patient or pooled patients together). The list of phylotypes found and their relationship to known bacterial taxa (using BLAST searches) was biologically informative. In addition to creating an inventory, the goals of the experiment were to better describe the bacteria communities and their differences along the intestinal tract or between patients. Only three (healthy) patients' biopsies were sequenced and classified in this way. This leads to a data set with 7 samples from each of the 3 patients. Clearly, with such a small sample size, the analysis cannot extrapolate to the population in general but can only focus on describing the patients observed.

2.2 Resulting Data

The end result of the experiment is several different types of data. At the most basic level, an observation corresponds to a sequenced DNA strand. Its data consists of two measurements: its 16S rDNA sequence and an indicator of which of the twenty-one samples it was found in. For the intestinal data, this led to $N = 11,831$ observations corresponding to the N strands of DNA sequenced. The DNA sequence, as we have mentioned, can be summarized in different ways, such as its similarity to other sequences, the phylotype to which it has been assigned, or its location in a

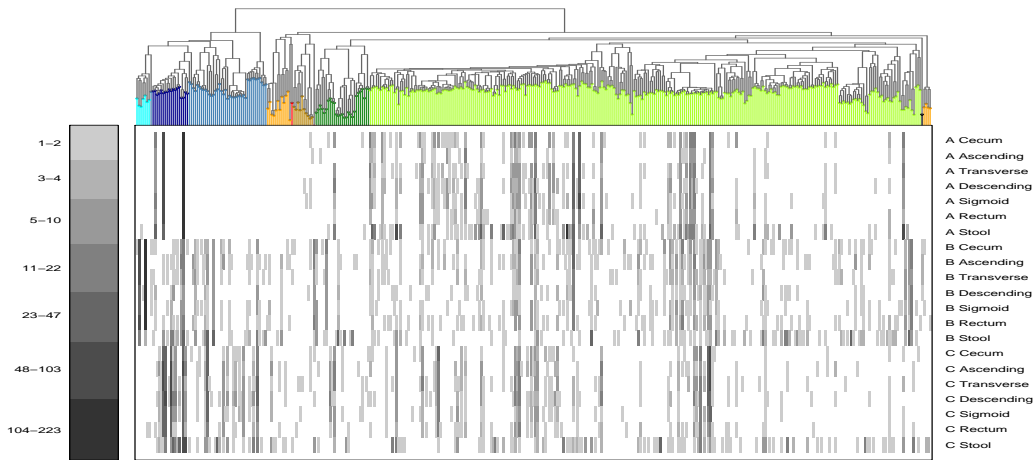


Figure 1: Depiction of the abundance matrix from Eckburg et al. (2005). Rows indicate samples, grouped by patient, and columns correspond to different phylotypes. The grey scale indicates the level of abundance on a log scale (see legend on left for conversion to original abundances). The colors on the phylogenetic tree indicate phylum, as in Eckburg et al. (2005), but with a different choice of colors: blue=*Bacteroidetes*, green=*Firmicutes*, yellow=various *Proteobacteria*, tan=*Verrucomicrobia*. We additionally colored two portions of the *Bacteroidetes* phylum (blue) separately: roughly identifiable as *Prevotallae* and *B. vulgatus* they are colored lightest blue and darkest blue, respectively. Also, we colored the *Firmicutes* (green) with two different shades for *B. Mollicutes* and *Clostridia* (dark green and light green, respectively).

phylogenetic tree built between different sequences. The analysis discussed in detail here will reduce the sequence data to the phylotype-level, ignoring the individual sequence data. This means that each of the N observations (or sequenced strands of DNA) belongs to one of $S = 395$ phylotypes. A phylogenetic tree for the phylotypes was built using maximum likelihood estimation of the tree (Felsenstein, 1981). Specifically, the tree was built using a representative instance of the 16S rDNA sequence from each phylotype, generally a consensus sequence of the sequences classified into that phylotype.

We can visualize both aspects of the data by juxtaposing the phylogenetic tree of the phylotypes with the numerical abundances of each phylotype across the samples, as in Figure 1. It is clear from the figure that there is a great deal of sparsity in the data; many phylotypes are present in low numbers and in only a few samples. At the same time, there are some highly abundant phylotypes found at high levels in most samples. From this visual inspection, we can also see the importance of jointly considering both aspects of the data – the abundances of the phylotypes as well as the relationships among the phylotypes. In particular, we see definite regions of the tree that seem to be dissimilar between the patients, such as the *Bacteroidetes* phylum (colored shades of blue) where patient A has much less abundance across all of his/her samples than the other two patients.

Given the large number of species (395) as compared to the number of samples (21), we could probably reorder the phylotypes and find other sets of phylotypes that are also very different across the patients. However, the clusters defined by the phylogenetic tree provide biological information separate from the numerical abundances regarding the relationships among the phylotypes; thus patterns of sparsity or differences amongst the patients following the clusters in the tree are generally of greater interest than an arbitrary grouping: there is known biological meaning to the group. The

additional information found by using the phylogenetic similarities can serve as a check on the kind of relationships among the phylotypes that we are interested in. This will be particularly important since we have so many more phylotypes than samples. Focusing the analysis to follow the structure of the tree will allow for more meaningful results.

Another important facet of the data that we can see in Figure 1 is the effect of using an arbitrary cutoff for defining phylotypes. The tree’s branch length reflects the similarity between the species. Some phylotypes clearly form tight bunches of very similar phylotypes, particularly in the *Clostridia* family of the *Firmicutes* phylum (light green). If we had changed the cutoff for defining phylotypes, we could imagine these groups collapsing into a few distinct phylotypes. Therefore, we need to be careful to have an analysis that is robust to such small changes and does not count each phylotype as equally important.

Remark 2.1. Note that we have estimated our tree from the same data (the 16S rDNA sequences) from which we created our table of abundances. In general trees that simply come from ordering the variables based on a data matrix – such as hierarchical clustering trees – will not provide separate, outside information about the data and are not appropriate for any of the analysis done here. However, in this instance, the abundances and the phylogenetic similarities are summaries of different aspects of the sequences we started with. As we described above, each observation contains data regarding its presence/absence in a sample and also its DNA sequence information. We could have a full tree from all of the DNA sequences where each leaf of the tree corresponds to just one DNA sequence; rows of the corresponding abundance matrix would be vectors of which sample the sequence was found. Reducing the tree to the phylotypes based on sequence similarity collapses rows of this expanded contingency table, a process not affected by the actual data in the expanded table but by the sequence similarities of the leaves of the expanded tree.

3 Generalized PCA (gPCA)

Before we describe DPCoA and the analysis of the bacterial data, we will describe the non-standard metrics that form the basis of the approach in this paper and also review a generalized form of PCA that incorporates these metrics. Note that the term ‘Generalized PCA’ is our terminology in order to simplify discussion. The method is often called the duality principle (Escoufier, 1987; Holmes, 2006; Dray and Dufour, 2007) and more generally it is called an “ordination” procedure (i.e. a method giving low-dimensional coordinates or ordinates to data). We will postpone giving examples of gPCA until section 4.2.

For vectors \mathbf{x} and \mathbf{y} in \mathbb{R}^p , let the positive definite matrix \mathbf{Q} define an inner product given by $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{Q}} = \mathbf{x}^T \mathbf{Q} \mathbf{y}$. Since \mathbf{Q} also defines a metric based on $\|\mathbf{x} - \mathbf{y}\|_{\mathbf{Q}}$, we may at times refer to \mathbf{Q} as a metric. A common example of such an inner-product is the Mahalanobis distance, where \mathbf{Q} is the inverse covariance matrix of the observed random vectors.

To use a non-standard metric space, we could simply transform the data by a set of linear combinations given by a decomposition of the metric and then perform any analysis of interest. However, some techniques are developed specifically with alternative metric spaces in mind. In particular, for PCA, there is a generalization of PCA common in ecological applications and categorical data applications that uses a wider choice of inner products spaces and metrics (Tenehaus and Young, 1985; Escoufier, 1987). The generalized PCA (gPCA) of the data matrix \mathbf{X} starts with the choice of a metric \mathbf{Q}_p in \mathbb{R}^p for the rows (observations) of \mathbf{X} and a metric \mathbf{Q}_n in \mathbb{R}^n for the columns of \mathbf{X} . These choices are abbreviated as the triplet $(\mathbf{X}, \mathbf{Q}_p, \mathbf{Q}_n)$ (see Escoufier (1987) for a more general explanation of the role of two separate metrics when viewing \mathbf{X} as an operator simultaneously in \mathbb{R}^p and \mathbb{R}^n).

In analogy with standard principal components, gPCA seeks the vector \mathbf{a} that maximizes $\text{var}(\langle \mathbf{a}, \mathbf{x} \rangle_{\mathbf{Q}_p})$, with \mathbf{a} constrained to have unit \mathbf{Q}_p -norm and successive \mathbf{a}_i constrained to be \mathbf{Q}_p -orthogonal to the preceding \mathbf{a}_i . As in PCA, the \mathbf{a}_i are eigenvectors, but now of the matrix $\Sigma \mathbf{Q}_p$ where Σ is the covariance matrix of \mathbf{x} . The matrix $\Sigma \mathbf{Q}_p$ is not symmetric, but because \mathbf{Q}_p is full rank, this is a well defined, positive definite generalized eigen equation, and the eigenvectors of $\Sigma \mathbf{Q}_p$ can be chosen to be a \mathbf{Q}_p -orthogonal set of vectors (see [Golub and van Loan \(1996\)](#)). The new coordinates for \mathbf{x} are given by $\mathbf{A}^T \mathbf{Q}_p \mathbf{x}$, where \mathbf{A} is the matrix with columns \mathbf{a}_i . As in PCA, we must estimate Σ from our data matrix \mathbf{X} ; we include the metric \mathbf{Q}_n for the columns in our estimate so that we have $\hat{\Sigma} = \mathbf{X}^T \mathbf{Q}_n \mathbf{X}$. Just as in PCA, gPCA best preserves inter-point similarities in the appropriate metric spaces.

A generalized form of the SVD of \mathbf{X} yields the solutions to gPCA on both the columns or the rows simultaneously. We can write $\mathbf{X} = \mathbf{B} \mathbf{\Lambda}^{1/2} \mathbf{A}^T$, where the columns of \mathbf{B} are \mathbf{Q}_n orthogonal and the columns of \mathbf{A} are \mathbf{Q}_p orthogonal. Then \mathbf{B} gives the solutions to the gPCA of the columns as observations while \mathbf{A} gives the solutions to the gPCA of the rows as observations. The corresponding eigen equations are

$$\begin{aligned} \mathbf{X}^T \mathbf{Q}_n \mathbf{X} \mathbf{Q}_p \mathbf{A} &= \mathbf{A} \mathbf{\Lambda} \\ \mathbf{X} \mathbf{Q}_p \mathbf{X}^T \mathbf{Q}_n \mathbf{B} &= \mathbf{B} \mathbf{\Lambda}, \end{aligned}$$

and $\mathbf{B} = \mathbf{X} \mathbf{Q}_p \mathbf{A} \mathbf{\Lambda}^{1/2}$.

4 DPCoA and Ecological Analyses of Species Abundance

We will now discuss the analysis of the bacterial data from [Eckburg et al. \(2005\)](#) using DPCoA. Species composition and comparison of species across different locations form the core of ecological studies. A large contingency table of species abundances for different locations is a common form of data in this literature; thus the analysis in [Eckburg et al. \(2005\)](#) naturally called for many techniques found in ecology. In particular, [Eckburg et al.](#) used a technique that, like PCA, gives a low dimensional representation of the data but incorporates the phylogenetic information. The technique, Double Principal Coordinate Analysis (DPCoA) ([Pavoine et al., 2004](#)), incorporates dissimilarity amongst the species directly into its representation of the locations and ultimately gives a representation of the locations that deals with the approximate species definition in the bacterial dataset. For this reason, [Eckburg et al.](#) relied on DPCoA to visualize both the relationships between the bacterial communities and the role of the phylogenetic relationships in this comparison.

We show that the technique can be described more simply as a gPCA, which gives avenues for generalization to other data settings. Thus, despite DPCoA's high specificity to ecological datasets in its original formulation, by reformulating the procedure we are able to show its relevancy in more general settings.

Our original interest was ecological, but large contingency tables appear in many other situations. For example, in document classification, the data could consist of the frequency of different words in different documents. Another example is allele frequency studies with the frequency of different alleles of a gene in different populations. We will continue to focus our notation and discussion on the phylogenetic/ecological scenario, but the methods presented here could be of use for these different data types.

4.1 Notation: the Weighted World of Abundance Tables

In this section, we will describe our notation as well as the standard transformations and terminology used in analyzing contingency tables.

Assume that the abundance of certain species are measured at L different locations and a total of S distinct species types are observed. Let \mathbf{A} be the resulting $L \times S$ contingency table of the observed abundances of species s at location ℓ . Because we are interested in comparing the species composition of the locations, we will represent each location by the *relative* proportion of the species in the location. A vector \mathbf{p}_ℓ of relative proportions at location ℓ is called a profile vector in the ecological literature and is obtained by dividing each row of \mathbf{A} by its row sum. The corresponding data matrix is given by $\mathbf{P} \in \mathbb{R}^{L \times S}$.

Let \mathbf{w}_L be a vector containing the row sums of \mathbf{A} normalized to sum to one, $\mathbf{w}_L = \mathbf{A}\mathbf{1}/N \in \mathbb{R}^L$. Then our matrix of location profiles, \mathbf{P} , is given by

$$\mathbf{P} = \begin{pmatrix} \mathbf{p}_1^T \\ \vdots \\ \mathbf{p}_L^T \end{pmatrix} = \mathbf{D}_{\mathbf{w}_L}^{-1} \mathbf{A}/N \in \mathbb{R}^{L \times S},$$

where $\mathbf{D}_{\mathbf{w}_L}$ is a diagonal matrix with diagonal elements given by \mathbf{w}_L respectively. The vector \mathbf{w}_L of relative row sums of \mathbf{A} also defines weights for each of the locations, and the weights are proportional to the total number of observations in that location. Since we would have lost this information by converting the data to relative frequencies, this information is preserved by using \mathbf{w}_L as weights for the observations (locations). The weighted mean of the locations, $\bar{\mathbf{p}}$ is given by $\mathbf{P}^T \mathbf{w}_L$ and the centered data matrix, $\tilde{\mathbf{P}}$, is given by $\tilde{\mathbf{P}} = (\mathbf{I} - \mathbf{1}\mathbf{w}_L^T)\mathbf{P}$.

4.2 A Few Important Properties of Contingency Tables

The Duality of Rows and Columns Note that the weighted mean, $\bar{\mathbf{p}}$, also sums to one and therefore is itself a potential location profile. In fact, $\bar{\mathbf{p}}$ is proportional to the column sums of \mathbf{A} , normalized to sum to one. Therefore, the weighted mean of the location profiles is the relative frequency of the species across *all* of the observations combined. If we had instead chosen to analyze the columns (species) as the observations, we would choose weights \mathbf{w}_S for the species in the same way as the rows: proportional to the total abundance in a species. Then we would have $\mathbf{w}_S = \bar{\mathbf{p}}$.

The equivalence of the column weights and the average row profile has interesting repercussions for data analysis because under this weighting scheme, we can equivalently center either the rows or the columns. If we let $\mathbf{M}_{\mathbf{w}_S} = (\mathbf{I}_S - \mathbf{1}_S \mathbf{w}_S^T)$ be the (weighted) centering matrix for the columns and $\mathbf{M}_{\mathbf{w}_L} = (\mathbf{I}_L - \mathbf{1}_L \mathbf{w}_L^T)$ be the (weighted) centering matrix of the rows, we have $\tilde{\mathbf{P}} = \mathbf{M}_{\mathbf{w}_L} \mathbf{P} = \mathbf{P} \mathbf{M}_{\mathbf{w}_S}$.

The role of variables in gPCA Because we analyze location profiles, there is a nice way to plot the variables (species) jointly with the observations (locations). Let \mathbf{e}_s be the standard basis vectors of \mathbb{R}^S . Then \mathbf{e}_s is also a profile vector representing a theoretical location that consists solely of species s . If we transform the data with an ordination technique, we can jointly transform \mathbf{e}_s and plot its transformation alongside the observed locations. Unlike the usual plots of variables, the coordinates of our rotated axes have a meaning as a data point and not just as a direction in space. So we can reasonably speak about distances between the location point to the phylotype point.

Examples of gPCA with Contingency Tables Since the data now consists of profiles such as \mathbf{p}_ℓ that are constrained to sum to one, different metric spaces are often used for analyzing contingency tables via gPCA. There is also usually additional information for either the rows or columns that we wish to use, traditionally the weights w_L and/or w_S . The most common example of gPCA is Correspondence Analysis (CA), which is a gPCA of the row profiles of a contingency table, and uses the triplet $(\tilde{\mathbf{P}}, \mathbf{D}_{w_S}^{-1}, \mathbf{D}_{w_L})$ (see [Greenacre, 1984](#), for a detailed treatment). This gives an inner product of the form $\mathbf{p}_k^T \mathbf{D}_{w_S}^{-1} \mathbf{p}_\ell$, down-weighting the more frequent species. This can be seen as counteracting a “size effect” for frequencies, where abundant species dominate the analysis; without this correction differences in rare species (which will be on a smaller order of magnitude) are lost.

However, one can argue that the weighting of CA places too much importance on low abundance species, even though those species are more likely to be miscounted and the corresponding data are probably less trustworthy. [Gimaret-Carpentier et al. \(1998\)](#) propose no weighting of the species, only the locations, which gives a triplet $(\tilde{\mathbf{P}}, \mathbf{I}_S, \mathbf{D}_{w_L})$ – just a regular PCA with weights on each observation. Such an analysis is also called Non-Symmetric Correspondence Analysis (NSCA).

4.3 DPCoA

DPCoA seeks to represent the relationship between the locations and species with meaningful measures of distance. In particular, given a pre-specified matrix Δ of dissimilarities between the species, the DPCoA gives new coordinates for the location profiles that have squared Euclidean distance between locations k and ℓ equal to

$$d(k, \ell) = (\mathbf{p}_k - \mathbf{p}_\ell)^T \left(-\frac{1}{2} \Delta\right) (\mathbf{p}_k - \mathbf{p}_\ell) = - \sum_{r \neq s} \Delta_{rs} (\mathbf{p}_{k(r)} - \mathbf{p}_{\ell(r)}) (\mathbf{p}_{k(s)} - \mathbf{p}_{\ell(s)}). \quad (1)$$

This distance is known as the Rao Dissimilarity ([Rao, 1982](#)). As a distance on the original profiles, it implies that differences between locations profiles are downweighted for the species that are similar and upweighted for very distinct species. We can see that for the bacterial example, this distance has the effect of not declaring samples very different if the differences occur in phylogenetically similar phylotypes.

As with PCA or Multi-Dimensional Scaling (MDS), restricting to successive dimensions of the coordinates from DPCoA conserves the most interpoint similarity, where the interpoint similarity is according to the distance given in equation (1). DPCoA also gives coordinates for the species that can be plotted along with the locations. Using these species coordinates, the squared Euclidean distance between species will be equal to their dissimilarities given in Δ . Additionally, the coordinates of the locations and species are aligned so that the coordinates of a location are equal to the weighted average of the coordinates of the species, where the weights are given by \mathbf{p}_ℓ , the profile for that location.

The technique of DPCoA is given by [Pavoine et al. \(2004\)](#) as a series of steps. We will see below (4.3.1) that we can more easily describe the method as a general kind of PCA; however, we will reiterate their original description of the technique here for comparison.

Assume that the *squared* pairwise distances/dissimilarities between the species are given by a $S \times S$ matrix Δ , or equivalently that the pairwise distances between the species are given in matrix, Υ . We also assume that Υ is Euclidean (i.e. coordinates can be found for the points so that the standard Euclidean distance between points are given by the entries of Υ).

1. Find Euclidean coordinates of the species using a weighted version of Multidimensional Scaling of Υ with weights for the species given by w_S , their relative abundance in all the samples.

Let these coordinates be given in the rows of the matrix $\mathbf{Z} \in \mathbb{R}^{S \times r}$ ($r \leq S-1$ is the dimension of the space required to contain the species)

2. Set the coordinates of the locations to be at the barycenter of the species coordinates. In other words, each location ℓ is given coordinates that are the weighted average of the coordinates of all the species and the weights are given by the relative abundance of the species in that site (which is contained in the vector \mathbf{p}_ℓ). Let the rows of the $L \times r$ matrix \mathbf{Y} contain the coordinates of the sites, so $\mathbf{Y} = \mathbf{PZ}$. The squared pairwise Euclidean distance between the locations using these coordinates will be equal to their Rao Dissimilarity using the dissimilarity matrix Δ .
3. Find a lower dimensional representation of the locations using a generalized principal components analysis on the triplet $(\mathbf{Y}, \mathbf{I}_S, \mathbf{D}_{w_L})$. These gives the new basis \mathbf{F} and then the final coordinates of the locations are given by $\mathbf{L} = \mathbf{YF}$. We also transform the coordinates of the species to get species coordinates, $\mathbf{K} = \mathbf{ZF}$ respectively.

Remark 4.1. The requirement that Υ be Euclidean not only makes the MDS in the first step of DPCoA possible, but also guarantees that the Rao Dissimilarity in equation (1) is a proper distance for vectors constrained to have the same, fixed, sum such as our profile vectors (see [Pavoine et al., 2004](#)).

4.3.1 DPCoA as a gPCA

We can actually describe DPCoA more simply as a gPCA with a metric based on a similarity matrix derived from Δ . If Δ is the matrix in DPCoA, a natural corresponding similarity matrix \mathbf{S}_x is given by

$$\mathbf{S}_x = \mathbf{1}_S \mathbf{x}^T + \mathbf{x} \mathbf{1}_S^T - \frac{1}{2} \Delta,$$

for some vector \mathbf{x} of positive values so that \mathbf{S}_x is positive definite. Then the resulting coordinates of the locations given by the *final* step of DPCoA are equivalent to the coordinates of a generalized PCA performed on the triplet $(\tilde{\mathbf{P}}, \mathbf{S}_x, \mathbf{D}_{w_L})$. The final coordinates of the species given in DPCoA can also be found from the gPCA on $\tilde{\mathbf{P}}$ and are simply the coordinates of the original axes of the *uncentered* data matrix \mathbf{P} , centered and then transformed by the transformation defined by the gPCA of $(\tilde{\mathbf{P}}, \mathbf{S}_x, \mathbf{D}_{w_L})$. This is independent of the choice of \mathbf{x} . See Appendix A for details.

Note that DPCoA and gPCA using \mathbf{S}_x are equivalent under the assumption that the input data matrix for the gPCA is $\tilde{\mathbf{P}}$, with the weighted centering described in section 4.1. If we chose any weight vector $\mathbf{w} \in \mathbb{R}^L$ to be weights for the locations, then we have that gPCA of $(\tilde{\mathbf{P}}, \mathbf{S}_x, \mathbf{w})$, where $\check{\mathbf{P}} = (\mathbf{I} - \mathbf{1} \mathbf{w}^T) \mathbf{P}$, is equivalent to DPCoA using the weights for the species given by $\mathbf{P}^T \mathbf{w}$, the weighted average of the species' abundance.

4.4 DPCoA Applied to Bacterial Data

The original analysis of the bacterial data relied on the technique of DPCoA reviewed above. This analysis used as the distance among the phylotypes their distance on the phylogenetic tree, Δ (the ‘‘patristic’’ distance, see below). We display in Figure 2 the ordination of the samples and species using the first two coordinates found by DPCoA using the implementation in the `ade4` package in R ([Chessel et al., 2005](#); [R Development Core Team, 2008](#)). The first obvious fact is that the patients are separated, almost entirely, by just the first axis. The first axis orders the patients B,C,A which correlates with visual examination of the data in Figure 1. Below we will compare DPCoA to other

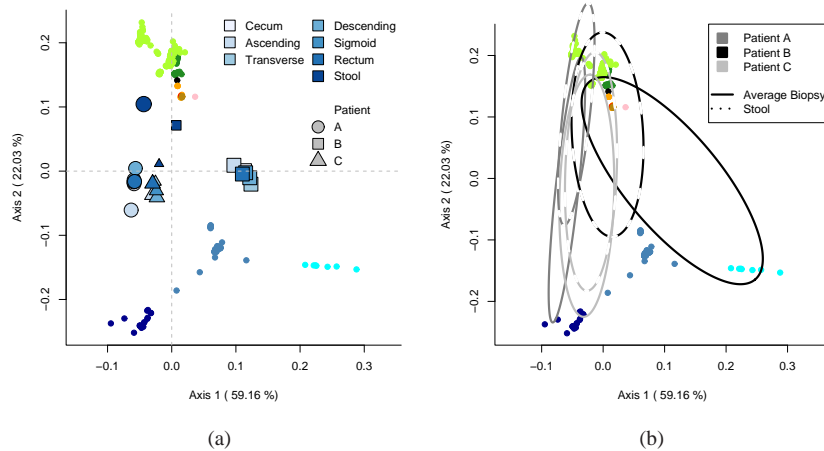


Figure 2: Scatter plot of the species and samples with the first two coordinates given by DPCoA. Species are shown as colored points in both plots. In plot (a), the samples are shown as the large blue shapes: different shapes indicate different patients and different shades of blue indicate location within the colon. In plot (b) samples are represented as ellipses that indicate the major directions of the abundances of the samples. For simplicity, a single ellipse for the combined abundance in the biopsies is shown because the internal biopsies are very similar.

ordination techniques and we will see that distinguishing the patients is not difficult since all of the techniques accomplish this, though not always in just one dimension. More interestingly, we also see in Figure 2 that the stool samples are distinguished from the internal biopsies of the colon, and the second axis seems to make this distinction. Again this makes sense from visually examining the data, since within each patient the stool samples do stand out from the biopsies.

The most striking aspect of the plot from DPCoA is the additional information provided from the inclusion of the phylotypes in the plot. Recall that when our data matrix is \mathbf{P} , our original axes e_s correspond to a location that is entirely concentrated in phylotype s . We showed that the coordinates of the phylotypes given by DPCoA will be the coordinates of our axis e_s centered and rotated like the observed profiles. Looking at the DPCoA plot, we see that the phylotypes' coordinates provide an interpretation for the first two dimensions. The phylotypes are in clusters much like the groupings on the tree – not surprising if we recall that in the full space the distances between the species are exactly the distances on the tree. What is interesting is how the clusters on the tree fill the space once projected into these two coordinates that preserve the Rao Dissimilarity among the locations. The distribution of the phylotypes indicate the importance of these clusters in determining the dissimilarity between the patients. Those far from the origin have more impact in defining the coordinates of the locations. We see the tension between the various *Bacteroides* (blue) and the rest of the tree.

Furthermore we can interpret the relationship between the locations and the phylotypes. We see that patient B is comparatively much more in the direction of the *Prevotallae*-like bacteria (light blue) while the other two patients are more in the direction of the *B. Vulgatus*-like phylotypes (dark blue). Similarly, the biopsies are comparatively more heavily represented in the *Bacteroides* (blue) portion of the tree while the stool samples are comparatively less so. Figure 2b depicts the different

samples as ellipses with the axes of the ellipses determined by the relative proportion of the different species for the location (plotting technique provided by `ade4`, see Supplementary Materials for details). This illustration emphasizes that the samples can be thought of giving weights to each phylotype, and the ellipse demonstrates the relative influence of the different species. We see graphically the different influences of the two groups of *Bacteriodes* (blue) in separating the biopsies of patient B from all of the rest of the samples. Transforming the data in various ways before analysis does not dramatically change these relationships (for example log-transforming the data or adding pseudo-counts).

All of these visualizations have, by necessity, focused on only the first two dimensions of the coordinates given by DPCoA. These dimensions do cover a large proportion of the Rao Dissimilarity, but still are only an approximation of the full space. Here we are interested in exploring the characteristics of the ordination procedure, but for permutation tests regarding the differences amongst the patients or between the biopsies and stool samples we would want to use the entire space.

4.4.1 Comparison to Other Approaches

How do these results compare to the other ordination techniques mentioned above? In Figure 3 we show the results of the ordination from Non-Symmetric Correspondence Analysis (NSCA), Correspondence Analysis (CA), and a Mahalanobis-like distance based on Σ^{-1} (see Section 6). We similarly center, rotate and project the axes e_s to get the species coordinates in the same manner as DPCoA.

As we mentioned, all of the techniques separate the three patients, but we see that DPCoA gives much more relevant results both in terms of the role of the species and in relating to our intuitive interpretation of the data. The NSCA (plot (b)) is the same technique as DPCoA but with each species at equal distance from every other; it is also just a standard PCA with weights on the observations. In the first two coordinates of the NSCA, we see that instead of having a smooth contribution from clusters of phylotypes, two individual phylotypes, far removed from the rest, contribute to the division of the patients much more than the rest. The bulk of the species have little contribution to these coordinates. Thus there is little from which to draw more general conclusions regarding the biological characteristics of the species which are influential. This is a consequence of treating each phylotype equally, rather than using the additional structure of the tree to shape the analysis. CA (plot (c)) on the other hand spreads out the importance of each phylotype. Here we can see the effect of the down-weighting metric in CA discussed earlier; differences found in the many low abundance phylotypes are allowed to influence the analysis. Rather than a couple of phylotypes dominating the analysis, as in NSCA, the phylotypes play more equal roles.

We might try to use any one of these techniques to reason out relationships amongst the variables. Each technique would give a different story in the role of the variables (phylotypes) dependent upon the assumptions inherent in the method. The relevant feature for our analysis is that we presuppose that a certain type of information is relevant – namely how the structure of the tree relates to the data. It focuses the analysis on finding an interpretation amongst the variables that follows the tree structure.

We note that the abundance table from metagenomic studies discussed here has many features common to high-throughput experiments in biology – in particular the number of biological samples is quite low compared to the number of measurements. We sought to integrate the phylogenetic information into the data analysis *a priori*. In this way, the analysis is constrained in a biologically relevant direction. In contrast, we could think of analyzing this abundance data much like a microarray experiment: test each phylotype individually for differences between the patients and use multiple testing criteria to identify individual phylotypes showing significant differences. A prob-

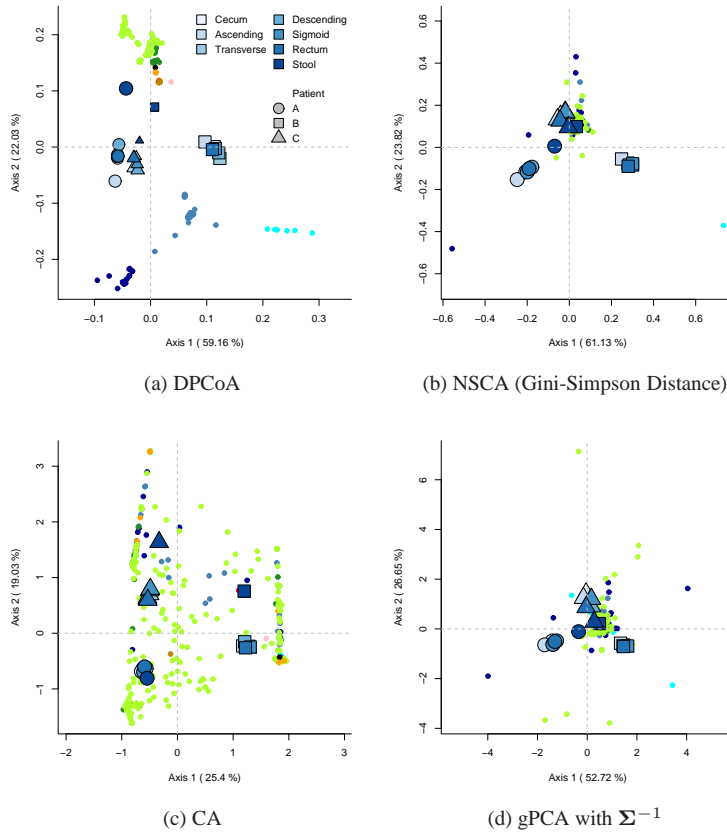


Figure 3: Coordinates of species and samples from alternative ordination techniques.

lem with this approach, which is also a common problem in microarrays analyses, would be that a list of significant phylotypes is difficult to interpret. In microarray studies, biological interpretation is often done *a posteriori* by then examining biological knowledge of the list of genes. We could similarly use the phylogenetic tree in this way. However, we just saw that an analysis independent of the tree highlighted only a couple of specific phylotypes from which it would be difficult to build a general connection to the tree.

5 Interpretation of Non-Standard Metrics

We see in the example of the bacterial data that a metric (or dissimilarity) that uses the phylogenetic information gives more biologically meaningful results. But what does the choice of a metric actually mean? Incorporating a metric for \mathbb{R}^p has an obvious rationale when the metric is a diagonal (implying weights for different variables) or when the metric is Σ^{-1} (Mahalanobis distance). However it is not immediately clear why a particular matrix \mathbf{Q}_p would improve a given data analysis.

Of course, a positive definite matrix \mathbf{Q}_p is always equivalent to some covariance matrix, and modeling the covariance between variables or observations is a simple way to create a candidate

matrix, Σ . But most statistical applications then choose Σ^{-1} to remove the covariance structure in the observations or variables. However, for our purposes, we want to highlight this information, not remove it, and then the appropriate choice is $\mathbf{Q}_p = \Sigma$. One intuitive rationale for this comes from thinking of the metric as defining a harmonic analysis of the data in the direction of the eigenvectors of \mathbf{Q}_p . This is also the perspective of [Rapaport et al. \(2007\)](#) in their proposal for the particular case of general graphs (see section 7).

More precisely, suppose \mathbf{Q}_p has an eigen decomposition given by $\mathbf{V}\Lambda\mathbf{V}^T$. Then each of p vectors \mathbf{v}_i define a linear combination of \mathbb{R}^p , and each $\tilde{x}_{(i)} = \mathbf{v}_i^T \mathbf{x}$ gives the magnitude of \mathbf{x} in the direction of the eigenvectors of \mathbf{Q}_p . Our original data \mathbf{x} can be written as

$$\mathbf{x} = \mathbf{V}\tilde{\mathbf{x}} = \sum_i \tilde{x}_{(i)} \mathbf{v}_i.$$

Depending on the nature of the \mathbf{v}_i , we can weight different directions to give more emphasis to these features. This gives a new vector $\mathbf{f}_w \in \mathbb{R}^p$,

$$\mathbf{f}_w(\mathbf{x}) = \sum_i w_{(i)} \tilde{x}_{(i)} \mathbf{v}_i.$$

A straightforward choice could be a simple thresholding function ($w_{(i)}$ is either 0 or 1) that would project \mathbf{x} onto the smaller subspace defined by the \mathbf{v}_i for which $w_{(i)} = 1$.

We can see that the standard inner-product between $\mathbf{f}_w(\mathbf{x})$ and $\mathbf{f}_w(\mathbf{y})$ is equivalent to taking the inner-product between \mathbf{x} and \mathbf{y} using the metric $\mathbf{V}\mathbf{D}_w^2\mathbf{V}^T$, where \mathbf{D}_w is the diagonal matrix with w as the diagonal:

$$\langle \mathbf{f}_w(\mathbf{x}), \mathbf{f}_w(\mathbf{y}) \rangle = \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{V}\mathbf{D}_w^2\mathbf{V}^T}.$$

Then the choice of a metric \mathbf{Q}_p is equivalent to the choice of weighting each \mathbf{v}_i by $\lambda_i^{1/2}(\mathbf{Q}_p)$ and $\mathbf{f}_w(\mathbf{x}) = \mathbf{Q}_p^{1/2} \mathbf{x}$.

A particular covariance matrix sets up a system of eigenvectors, which can then be weighted in different ways to create a family of metrics. Σ and Σ^{-1} obviously have the same eigenvectors and differ in the weighting the eigenvectors: Σ places emphasis on the directions with more information about the outside structure, while Σ^{-1} emphasizes directions with the least information about the outside structure. Depending on whether this structure is thought to enlighten or confound the analysis, the different weighting systems are used.

Such a weighted transformation is, of course, analogous to harmonic analysis or wavelet analysis for functional data. PCA could also be described similarly, only with the \mathbf{v}_i dependent on the observed variability of the data. In these cases, the basis functions can be ordered to hopefully reflect increasingly less meaningful variations of the data, so that the important information in the data is captured in the first few directions. Eigenvectors of a general covariance matrix describe linear combinations of decreasing variance, and thus presumably decreasing ability to reveal the structure of interest. This is of course the rationale in being able to order and weight the eigenvectors in importance in more general settings.

A change of basis is particularly useful in the analysis of the data if the bases are 'sparse' – non-zero in a small portion of the coordinate space; the resulting transformation of the data is easily interpreted as contrasts or combinations of a small set of variables. This is the appeal of wavelets or various sparse PCA algorithms.

6 A Metric for Species Related by a Phylogenetic Tree

There is a natural model that gives a covariance model for the leaves on the tree (the existing species). We will see that the eigenvectors of this covariance matrix also have nice localization properties relative to the tree and also ultimately relate back to our DPCoA of the bacterial data.

6.1 A Covariance Model for Related Species

In comparing data between species, it is assumed that measurements for closely related species will be similar and that some of their similarity will be due solely to their common ancestor (as opposed to their similar external constraints, for example). There is a common model to describe the dependence expected solely due to evolution; it is based on representing how and when the trait evolved over time, given from the phylogenetic tree of how the ancestral species diverged.

Assume that we have a known phylogenetic tree describing the ancestral relationship of S extant species and a trait of interest for these species that has evolved over time according to the evolution depicted on this tree. The most common probabilistic model for the evolution of the value of this trait, due to [Cavalli-Sforza and Piazza \(1975\)](#), is one of Brownian motion model over time (but see [Hansen and Martins \(1996\)](#) for other plausible alternatives); furthermore at each speciation event, the evolution model assumes that the resulting sister species continue to evolve independently. At our current time, we observe S species and measure the trait for these species. This gives data values $y_{(s)}$ for each of the existing species. In fact, we generalize and assume for each of the S species, we observe the trait at possibly different times $t_{(s)}$.

Under this model, the final observed values of the S species on the tree will of course follow a multivariate normal distribution. The covariance between species r and s will be proportional to the total length of time that the processes of the two species were identical – the amount of history that the two species shared until their lineages diverged. In other words, $\text{cov}(y_{(s)}, y_{(r)}) = \sigma^2 t_{rs}$, where t_{rs} is the time at which the two lineages diverged, as measured from their most common ancestor (we have conditioned on the value of the common ancestor of all the S species).

We can write this covariance quite simply in terms of the topology of the tree and the length of the branches, assuming that the branch length is reflective of evolutionary time. Let $d_{\mathcal{T}}(\cdot, \cdot)$ be defined as the length of the shortest path between any two nodes of the tree (the patristic distance), \mathcal{R} be the ancestor at the root of the tree (the most recent common ancestor between all the S species), and Δ be the distance matrix of the leaves based on the distance $d_{\mathcal{T}}(\cdot, \cdot)$. Then we can write the covariance matrix Σ as

$$\Sigma_{rs} = 1/2(d_{\mathcal{T}}(r, \mathcal{R}) + d_{\mathcal{T}}(s, \mathcal{R}) - \Delta_{rs}).$$

Connection with DPCoA Note that $\Sigma = 1/2(\mathbf{1}\mathbf{t}^T + \mathbf{t}\mathbf{1}^T - \Delta)$, where \mathbf{t} is the vector of the observation time of each species. We see that Σ is a similarity matrix as described in section 4.3.1. Therefore, DPCoA on profile data using the patristic distance between species is equivalent to using gPCA with Σ as the species metric.

Σ in Phylogenetics This model of evolution is a fundamental element of the most common method used in phylogenetics for analyzing what is called “comparative data” – data where the observational units are different species. Comparative methods model relationships between different continuous traits where the observations are measurements of the traits on different species. There is obvious correlation in the observations (species) due to evolutionary chance and not due to any necessary relationship between traits. From a purely statistical point of view and with the

the model given above, this dependency among the observations can easily be rectified using Generalized Least Squares – regression based on the metric induced by Σ^{-1} . This is the proposal of Grafen (called Phylogenetic GLS). The popular Phylogenetic Independent Contrasts (PIC) method of Felsenstein (1985) is also equivalent to GLS (Purdom, 2006).

6.2 Properties of Phylogenetic Metric

There is an obvious hierarchy for the phylogenetic tree which we hope the metric reflects. We would like that the eigenvectors of Σ be sparse in a useful way relative to the structure of the tree, for example that they contrast sister subtrees of the phylogenetic tree and be zero elsewhere. Furthermore, we would like that eigenvectors give increasingly specific level of detail so that eigenvectors corresponding to larger eigenvalues highlight deeper structure in the tree. Put together, these statements would imply that the eigenvectors offer a multi-scale analysis of the tree, with eigenvectors corresponding to large eigenvalues interpretable as summarizing differences in the large initial partitions of the tree and smaller eigenvalues giving eigenvectors reflecting the distinctions between the later divisions of the tree.

Several authors in phylogenetics have asserted that the eigenvectors of Σ have this multi-scale structure (e.g. Cavalli-Sforza and Piazza, 1975; Rohlf, 2001; Martins and Housworth, 2002). Only limited statements of this kind can be rigorously made about a phylogenetic tree with more than four leaves/species (see Purdom (2006) for a longer discussion). But empirical observations of the eigenvectors show that they often do often have some characteristics of this multi-scale property. Because Σ has a block structure, we are automatically guaranteed that the eigenvectors of Σ will, at a minimum, be non-zero for only one side or the other of the initial split in the tree (see Appendix C for more). Beyond this, if we ignore the comparatively small values in the eigenvector, eigenvectors corresponding to smaller eigenvalues do tend to divide the species into smaller and smaller groups based on the sign of the entries. Examining the eigenvectors of Σ for the phylogenetic tree from the bacterial data example (see Supplementary Figures for plots of a random sample of eigenvectors), we see that the eigenvectors are composed of a few well defined groups of species (again ignoring small elements of the eigenvectors). Though the groupings within an eigenvector do not just correspond to sister subtrees, the groups do tend to correspond to more closely related subtrees.

6.3 Effect of the Choice of Metric

We can see the effect of using Σ in our gPCA by examining the linear combinations that gPCA using Σ chooses. For any ordination technique, let \mathbf{V} be a matrix that rotates the *original* profiles \mathbf{P} to give us the final ordination; in gPCA of centered data, this will be the matrix $\mathbf{M}_{w_S} \mathbf{Q}_S \mathbf{A}$. We examine the different linear transformations, v_i , from gPCA with Σ as compared to the transformation for a standard PCA on the data $\tilde{\mathbf{P}}$ (equivalently, NSCA). And we also compare to the eigenvectors of Σ : if the covariance between the species was exactly the Σ predicted by the evolution model, then these would be the principal components of such data. Thus we can think of the eigenvectors of Σ as PCA on the tree.

In Figure 4 we order the elements of v_i from these three ordination techniques so that they line up with the phylogenetic tree. In this way we can see the relative importance of the phylotypes in transforming the data. When we look at the linear combinations for the first few coordinates, we see that the principal components from our gPCA with Σ intuitively seem to be a trade off between these two options, and we could think of this as a shrinking of the data variability in the “direction” of the tree. This is a particularly appealing idea, since we are treating the phylotypes as variables and there are far too many variables for the number of samples we have.

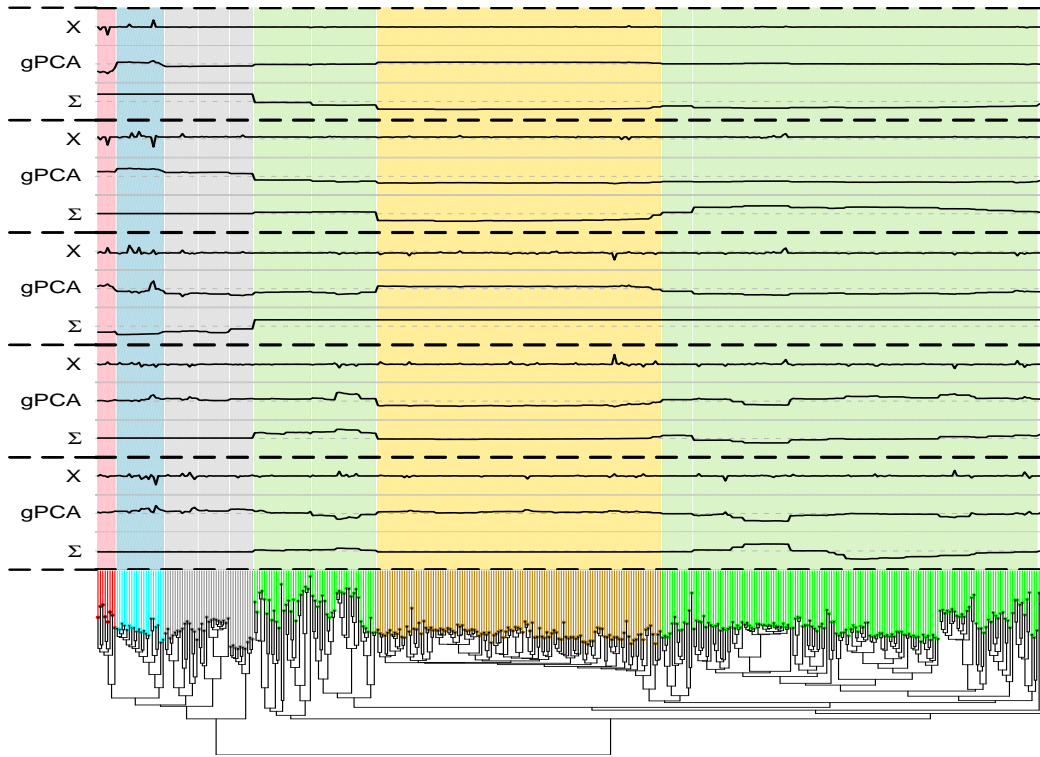


Figure 4: Shown are the first five linear combinations of gPCA using Σ that act on the observations in \mathbf{P} (the location profiles) to create the first five coordinates (v_i). The five dimensions are divided by thick, dotted line. Also shown adjacent to each gPCA vector are the linear combinations from a standard PCA of $\tilde{\mathbf{P}}$ (labeled ‘X’) and the eigenvectors of Σ (labeled ‘ Σ ’).

Despite the intuitive results, the analysis depends on our choice of encoding the tree using Σ (or equivalently, for DPCoA, our choice of Δ). In particular, the block structure of Σ puts large emphasis on the first initial partition of the species at the root of the tree; these two groups of species are considered independent, conditional on the root ancestor. We can see this emphasis on this first divide from the Rao Dissimilarity based on Δ , where these two lineages will be far away from each other and thus differences between will be accorded more weight in the analysis. Thus, the analysis is sensitive to the definition of the root as well as to changes deep within the tree.

However, changes near the tips of the tree, both in the numerical data and the definition of the tree, will have little impact on the gPCA. For the bacterial data that we are interested in, the deeper tree structure is more trustworthy than the structure near the leaves of the tree because of the approximate definition of species. It is a reasonable compromise to put more weight on the deeper structure of the tree, and base our analysis on this dependence, in exchange for resolving the more fundamental problem in our definition of the species.

We can see some effect of the choice of root in the tree in the bacterial data where the two lineages are basically the *Bacteriodes* and the *Firmicutes*; this break is strongly seen in Figure 2, particularly in the second axis. However, the first axis was dominated more by differences within

the *Bacteriodes*. Thus, as illustrated by the comparison of the v_i in Figure 4, we see a compromise between the patterns that occur only the data and those in Σ .

7 General Graphs

It is clear that the same approach is applicable to other situations where there is complicated information that is related to the experimental data. By understanding our phylogenetic analysis as a specific example in a general approach to data analysis, we can compare with other techniques as well as take advantage of insights from other data situations.

A closely related example is when we have not a phylogenetic tree, but a more general graph structure that describes the relationship of our variables or observations. The analysis of experimental data in tandem with related biological networks by Rapaport et al. (2007) is equivalent to our metric approach. There, the authors used the Laplacian matrix associated with a graph to represent the biological graphs that related genes. The Laplacian matrix is a natural choice for graphs; the eigenvectors have similar multiscale properties as our metric for the phylogenetic tree. In appendix B we briefly discuss the possibility of treating the phylogenetic tree as a general graph and using the Laplacian as a metric. We chose another approach here because such a choice does not well reflect the phylogenetic information in the tree.

8 Conclusion

There is a clear necessity for including phylogenetic information in an analysis of metagenomic data. gPCA gives a simple and compelling way to accomplish this. We also see from our recasting of DPCoA as a gPCA, that the framework of gPCA allows for easy comparisons between analyses as well as further exploration as to the effect of our choice of metrics.

The use of non-standard metrics is quite natural in statistics and can be implemented in a variety of ways, PCA being merely the simplest. Common examples, such as Mahalanobis distance, are usually data-driven, but we see that metrics based on outside knowledge can be used to include complicated and heterogeneous information into the analysis of our numerical data. This kind of information can help to give more context to the data, particularly when the number of variables is large as compared to the samples. Moreover, since the metrics here correspond to covariance matrices, probabilistic models give a simple approach for encoding information appropriately. Often, as in the case of phylogenetic trees, the eigenvectors of such covariance matrices have nice localization properties that highlight the relevant spatial or regional patterns of the prior information.

9 Acknowledgements

This work was supported in part by a NSF Post-doctoral Fellowship in Bioinformatics and by NSF Grant 02-41246.

References

BAPAT, R., KIRKLAND, S. and NEUMANN, M. (2005). On distance matrices and Laplacians. *Linear Algebra and its Applications* **401** 193–209.

- CAVALLI-SFORZA, L. and PIAZZA, A. (1975). Analysis of evolution: Evolutionary rates, independence and treeness. *Theoretical Population Biology* **8** 127–165.
- CHESSEL, D., DUFOUR, A.-B., DRAY, S., WITH CONTRIBUTIONS FROM JEAN R. LOBRY, OLLIER, S., PAVOINE, S. and THIOULOUSE., J. (2005). *ade4: Analysis of Environmental Data : Exploratory and Euclidean methods in Environmental sciences*. R package version 1.4-1.
URL <http://pbil.univ-lyon1.fr/ADE-4>
- COMMITTEE ON METAGENOMICS (2007). *THE NEW SCIENCE OF METAGENOMICS: Revealing the Secrets of Our Microbial Planet*. THE NATIONAL ACADEMIES PRESS.
- DIESTEL, R. (2005). *Graph Theory*. 3rd ed. Graduate Texts in Mathematics, Springer, New York.
- DRAY, S. and DUFOUR, A.-B. (2007). The ade4 package: Implementing the duality diagram for ecologists. *Journal of Statistical Software* **22**.
URL <http://www.jstatsoft.org/v22/i04>
- ECKBURG, P. B., BIK, E. M., BERNSTEIN, C. N., PURDOM, E., DETHLEFSEN, L., SARGENT, M., GILL, S. R., NELSON, K. E. and RELMAN, D. A. (2005). Diversity of the human intestinal microbial flora. *Science* **308** 1635–1638.
- ESCOUFIER, Y. (1987). The duality diagram: A means for better practical applications. In *Developments in Numerical Ecology* (P. Legendre and L. L., eds.), vol. G14 of *NATO ASI Series*. Springer-Verlag, Berlin, 139–156.
- FELSENSTEIN, J. (1981). Evolutionary trees from gene frequencies and quantitative characters: Finding maximum likelihood estimates. *Evolution* **35** 1229–1242.
- FELSENSTEIN, J. (1985). Phylogenies and the comparative method. *American Naturalist* **125** 1–15.
- GIMARET-CARPENTIER, C., CHESSEL, D. and PASCAL, J. P. (1998). Non-symmetric correspondence analysis: an alternative for community analysis with species occurrences data. *Plant Ecology* **138** 97–112.
- GOLUB, G. H. and VAN LOAN, C. F. (1996). *Matrix Computations*. 3rd ed. The Johns Hopkins University Press, Baltimore.
- GRAFEN, A. (1989). The phylogenetic regression. *Phil. Trans. Royal Society of London, Series B* **326** 119–157.
- GREENACRE, M. J. (1984). *Theory and applications of correspondence analysis*. Academic Press, Orlando, Florida.
- HANSEN, T. F. and MARTINS, E. P. (1996). Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific data. *Evolution* **50** 1404–1417.
- HOLMES, S. (2006). Multivariate analysis: The french way. In *Festschrift for David Freedman*. IMS Lecture Notes, IMS.
- KONDOR, R. I. and LAFFERTY, J. (2002). Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of ICML*.

- MARTINS, E. P. and HOUSWORTH, E. A. (2002). Phylogeny shape and the phylogenetic comparative method. *Syst. Biol* **51** 873–880.
- PAVOINE, S., DUFOUR, A.-B. and CHESSEL, D. (2004). From dissimilarities among species to dissimilarities among sites: a double principal coordinate analysis. *Journal of Theoretical Biology* **228** 523–537.
- PURDOM, E. (2006). *Multivariate Kernel Methods in the Analysis of Graphical Structures*. Ph.D. thesis, Stanford University.
- R DEVELOPMENT CORE TEAM (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL <http://www.R-project.org>
- RAO, C. R. (1982). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology* **21** 24–43.
- RAPAPORT, F., ZINOVYEV, A., DUTREIX, M., BARILLOT, E. and VERT, J.-P. (2007). Classification of microarray data using gene networks. *BMC Bioinformatics* **8**.
- ROHLF, F. J. (2001). Comparative methods for the analysis of continuous variables: Geometric interpretations. *Evolution* **55** 2143–2160.
- TENEHAUS, M. and YOUNG, F. W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis, and other methods for quantifying categorical multivariate data. *Psychometrika* **50** 91–119.

A DPCoA and gPCA

Lemma. Let Δ be given as distances between the species and let $\mathbf{S}_x = \mathbf{1}x^T + x\mathbf{1}^T - \frac{1}{2}\Delta$ where x is such that \mathbf{S}_x is positive definite. Then the coordinates for the locations given by \mathbf{L} in DPCoA using Δ is equivalent to the coordinates $\hat{\mathbf{X}}$ of the locations given by gPCA with the triplet $(\tilde{\mathbf{P}}, \mathbf{S}_x, \mathbf{D}_{w_L})$. Furthermore, the coordinates of the species given by DPCoA in the matrix \mathbf{K} are equivalent to the coordinates obtained by centering and then rotating the original axes e_s by the transformation implied from the gPCA of $(\tilde{\mathbf{P}}, \mathbf{S}_x, \mathbf{D}_{w_L})$ so that $\mathbf{K} = \mathbf{M}_{w_S} \mathbf{S}_x \mathbf{A}$

Proof. Recall that $\mathbf{M}_{w_S} = (\mathbf{I}_S - \mathbf{1}_S w_S^T)$ is a weighted centering matrix for a set of S observations with weights given by w_S .

We examine the resulting coordinates of the two methods. We recall that the weighted MDS of the pairwise species matrix gives species coordinates $\mathbf{Z} = \mathbf{D}_{w_S}^{-1/2} \mathbf{U} \mathbf{\Lambda}^{1/2}$ where \mathbf{U} are the eigenvectors of $\mathbf{D}_{w_S}^{1/2} \mathbf{M}_{w_S} (-\Delta/2) \mathbf{M}_{w_S}^T \mathbf{D}_{w_S}^{1/2}$. Then final coordinates \mathbf{L} of the locations from DPCoA are given from the decomposition of the triplet $(\mathbf{Y}, \mathbf{I}_S, \mathbf{D}_{w_L})$, where \mathbf{Y} holds the coordinates of the locations obtained by taking the barycenter of the species coordinates using the location profiles as weights; then $\mathbf{Y} = \mathbf{PZ} = \mathbf{PD}_{w_S}^{-1/2} \mathbf{U} \mathbf{\Lambda}^{1/2}$. The fundamental set of eigen equations for this final gPCA are

$$\begin{aligned} \mathbf{YD}_{w_L} \mathbf{Y}^T \mathbf{F} &= \mathbf{F} \Phi & \text{where} & & \mathbf{F}^T \mathbf{F} &= \mathbf{I}_r \\ \mathbf{Y} \mathbf{Y}^T \mathbf{D}_{w_L} \mathbf{G} &= \mathbf{G} \Phi & & & \mathbf{G}^T \mathbf{D}_{w_L} \mathbf{G} &= \mathbf{I}_r \end{aligned} \quad (2)$$

and $\mathbf{Y} = \mathbf{G}\Phi^{1/2}\mathbf{F}^T$ is the generalized SVD decomposition of \mathbf{Y} .

The fundamental eigen equations for a gPCA of the triplet $(\tilde{\mathbf{P}}, \mathbf{S}_x, \mathbf{D}_{w_L})$, on the other hand, are

$$\begin{aligned} \tilde{\mathbf{P}}^T \mathbf{D}_{w_L} \tilde{\mathbf{P}} \mathbf{S}_x \mathbf{A} &= \mathbf{A} \Psi & \text{where} & & \mathbf{A}^T \mathbf{S}_x \mathbf{A} &= \mathbf{I}_r \\ \tilde{\mathbf{P}} \mathbf{S}_x \tilde{\mathbf{P}}^T \mathbf{D}_{w_L} \mathbf{B} &= \mathbf{B} \Psi & & & \mathbf{B}^T \mathbf{D}_{w_L} \mathbf{B} &= \mathbf{I}_r, \end{aligned} \quad (3)$$

so that $\mathbf{P}\mathbf{M}_{w_S} = \mathbf{B}\Psi^{1/2}\mathbf{A}^T$ is the corresponding gSVD.

Note that $\mathbf{M}_w \mathbf{S}_x \mathbf{M}_w^T = \mathbf{M}_w (-\frac{1}{2}\Delta) \mathbf{M}_w^T$ for any choice of weights w and vector x . Since $\tilde{\mathbf{P}} = \mathbf{M}_{w_L} \mathbf{P} = \mathbf{P}\mathbf{M}_{w_S}$ and $\mathbf{Y} = \mathbf{w}$ we see that \mathbf{B} and \mathbf{G} are both eigenvectors for the same matrix, $\mathbf{P}\mathbf{M}_{w_S} \Delta \mathbf{M}_{w_S}^T \mathbf{P}^T \mathbf{D}_{w_L}$, implying that \mathbf{B} and \mathbf{G} are the \mathbf{D}_{w_L} -orthonormal eigenvectors of the same matrix. This implies that the eigenvalues are the same ($\Phi = \Psi$) and that \mathbf{B} and \mathbf{G} are the same up to a sign change (assuming unique eigenvalues).

The resulting coordinates for the locations under DPCoA are given by $\mathbf{L} = \mathbf{Y}\mathbf{F} = \mathbf{G}\Phi^{1/2}$. With gPCA of $(\tilde{\mathbf{P}}, \mathbf{S}_x, \mathbf{D}_{w_L})$, the location coordinates are $\hat{\mathbf{X}} = \mathbf{P}\mathbf{M}_{w_S} \mathbf{S}_x \mathbf{A} = \mathbf{B}\Psi^{1/2}$ and therefore we have that $\mathbf{L} = \hat{\mathbf{X}}$ – the coordinates of the locations are the same in the two methods.

The coordinates for the species are given by DPCoA as the rotation of the coordinates given in \mathbf{Z} by \mathbf{F} : $\mathbf{K} = \mathbf{Z}\mathbf{F}$. By the gSVD decomposition of \mathbf{Y} , we can write $\mathbf{F}^T = \Phi^{-1/2} \mathbf{G}^T \mathbf{D}_{w_L} \mathbf{Y}$ and similarly $\mathbf{B}\Psi^{-1/2} = \mathbf{P}\mathbf{M}_{w_S} \mathbf{S}_x \mathbf{A} \Psi^{-1}$. Remembering that $\mathbf{Z}\mathbf{Z}^T = \mathbf{M}_{w_S} \Delta \mathbf{M}_{w_S}^T$, the final coordinates of the species from DPCoA are given by

$$\begin{aligned} \mathbf{K} &= \mathbf{Z}\mathbf{Y}^T \mathbf{D}_{w_L} \mathbf{G} \Phi^{-1/2} = \mathbf{Z}\mathbf{Z}^T \mathbf{P}^T \mathbf{D}_{w_L} \mathbf{G} \Phi^{-1/2} \\ &= \mathbf{M}_{w_S} \Delta \mathbf{M}_{w_S}^T \mathbf{P}^T \mathbf{D}_{w_L} \mathbf{G} \Phi^{-1/2} \\ &= \mathbf{M}_{w_S} \mathbf{S}_x \underbrace{\mathbf{M}_{w_S}^T \mathbf{P}^T \mathbf{D}_{w_L} \mathbf{P} \mathbf{M}_{w_S} \mathbf{S}_x \mathbf{A}}_{=\mathbf{A}\Psi \text{ from (3)}} \Psi^{-1} \\ &= \mathbf{M}_{w_S} \mathbf{S}_x \mathbf{A} \end{aligned}$$

upto the sign change difference between \mathbf{G} and \mathbf{B} . ■

B The Laplacian and a Laplacian for Trees

The Laplacian matrix that is associated with the graph is given by $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where \mathbf{A} is the adjacency matrix of the graph and \mathbf{D} is the diagonal matrix consisting of the degree of each vertex. The spectral decomposition of \mathbf{L} is closely related to certain properties of the graph; in particular, there are many results linking the eigenvalues of \mathbf{L} with fundamental characteristics of the graph (see [Diestel, 2005](#)). There are fewer explicit characterizations of the eigenvectors that hold for all graphs. In a general way, the eigenvectors corresponding to small eigenvalues of \mathbf{L} represent large divisions in the graph (indeed for $\lambda_0 = 0$, we have the eigenvector $\mathbf{1}$ which is an average of all the nodes); they tend to be zero for large portions of the graph and the non-zero components are the same sign distinct regions of the graph. Those eigenvectors corresponding to large eigenvalues tend to be dominated by linear combinations of ‘close’ nodes or smaller groups of nodes and represent the ‘noisy’, small differences within neighboring vertices. Thus the eigenvectors of the Laplacian have ‘multiscale’ characteristics, particularly those eigenvectors corresponding to the largest and smallest of the eigenvalues. For data x associated with a graph, with each element of x corresponding to a node in the graph, the metrics for a graph based on the Laplacian will usually put greater weight on the eigenvectors corresponding to small eigenvalues, for example $1/\lambda_i$ or $\exp(-1/\lambda_i)$. This choice corresponds to the behavior of the eigenvectors.

The Laplacian gives the covariance between nodes from a useful model for describing relationships among the nodes – a model of diffusion of information through the graph. The covariance from this model is given by $\exp(-2\alpha\mathbf{L})$, known as the *heat kernel* of the graph (see [Kondor and Lafferty, 2002](#), for review). Of course this is equivalent to weighting the eigenvectors of the Laplacian with weight function $\exp(-\alpha\lambda_i)$.

A phylogenetic tree is, of course, a graph, and the Laplacian of a tree and the distances between nodes on a tree are quite simply related ([Bapat et al., 2005](#)). Let Δ_T be the distance matrix of the patristic distances between *all* the nodes of the tree (internal nodes as well as the leaves), and let \mathbf{L} be the Laplacian of the tree with weights $1/d(r, s)$ on each edge. Then we have that

$$\mathbf{L} = \mathbf{v}\mathbf{v}^T / \sum d(r, s) - 2\Delta_T^{-1},$$

where for a phylogenetic tree v is -1 or 1 depending on whether the node is a leaf of the tree or not.

However, since our data is observed on only certain nodes of the graph – the leaves of the tree – we need a metric that gives a relationship only between the leaves. If we use the Laplacian as our phylogenetic metric, we would have to constrain ourselves to the portion of the metric that corresponds to the relationships between just the leaves, \mathbf{L}_S . If we took as our metric the inverse of the Laplacian – which corresponds to an appropriate ordering of the eigenvectors by weighting each by $1/\lambda_i$ – we have that \mathbf{L}_S^{-1} is given by

$$\mathbf{L}_S^{-1} = c\gamma\gamma^T - 1/2\Delta_S,$$

where

$$c = (8\mathbf{1}^T \Delta_{S \times I} \mathbf{1})^{-1}$$

$$\gamma = \Delta_T \mathbf{v},$$

and $\Delta_S \in \mathbb{R}^{S \times S}$ is the distance matrix restricted to the distances between leaves of the tree and $\Delta_{SI} \in \mathbb{R}^{S \times S-1}$ is the distance matrix restricted to the distances between the leaves of the tree and $S-1$ internal nodes of the tree. This is an expression somewhat similar to our similarity matrix for DPCoA, but note that a gPCA based on \mathbf{L}_S^{-1} is not equivalent to DPCoA because $\mathbf{M}\gamma\gamma^T\mathbf{M}^T$ does not vanish.

However, restricting the metric to those portions dealing only with the leaves makes the metric difficult to interpret. The Laplacian restricted to the leaves will no longer have the same eigenvectors as the Laplacian and thus loses its connection to the behavior shown by the eigenvectors of the Laplacian. Furthermore, from the point of view of covariance modeling, the phylogenetic tree represents an evolutionary story that is more directly modeled by Σ .

C Eigenvectors of Σ for a Phylogenetic Tree

Note the block structure in Σ : if the root ancestor, \mathcal{R} , has immediate descendants \mathcal{P}_1 and \mathcal{P}_2 , then the covariance between any of the existing descendants of \mathcal{P}_1 and those of \mathcal{P}_2 will be 0. Thus we can order the rows and columns of Σ so that

$$\Sigma = \begin{pmatrix} \Sigma_1 & \emptyset \\ \emptyset & \Sigma_2 \end{pmatrix} \quad (4)$$

where Σ_1 is a $S_1 \times S_1$ matrix, S_1 is the number of extant species descended from \mathcal{P}_1 , and similarly with Σ_2 . This means that the eigenvectors of Σ must be of the form

$$\begin{pmatrix} \mathbf{v}_{1i} \\ \emptyset \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} \emptyset \\ \mathbf{v}_{2j} \end{pmatrix} \quad (5)$$

where $\{v_{1i}\}_{i=1}^{S_1}$ are the eigenvectors of Σ_1 and $\{v_{2j}\}_{j=1}^{S_2}$ are the eigenvectors of Σ_2 . Therefore, every eigenvector of Σ , at a minimum, must be only non-zero for one of the lineages.

Indeed, if we think back to the definition of Σ , the elements of the blocks Σ_1, Σ_2 are themselves rank-1 perturbations of block diagonal matrices:

$$\Sigma_1 = \begin{pmatrix} \Sigma_{11} & \emptyset \\ \emptyset & \Sigma_{12} \end{pmatrix} + c_1 \mathbf{1}\mathbf{1}^T \quad \Sigma_2 = \begin{pmatrix} \Sigma_{21} & \emptyset \\ \emptyset & \Sigma_{22} \end{pmatrix} + c_2 \mathbf{1}\mathbf{1}^T \quad (6)$$

where $c_1 = d_T(\mathcal{R}, \mathcal{P}_1)$ and $c_2 = d_T(\mathcal{R}, \mathcal{P}_2)$ (here we have assumed that $t \propto \mathbf{1}$ for simplicity). This same logic continues so that each sub-block can be written as a block matrix plus a rank-one perturbation. Σ thus consists of such nested rank-1 perturbations of block matrices.

The claims in the literature for a relationship of the eigenvectors of Σ to the partitions of the tree all stem from the comments of [Cavalli-Sforza and Piazza \(1975\)](#). They make assertions which they prove only in the case of a tree with four leaves ($S = 4$) and under the assumption of a constant rate of evolution ($t \propto \mathbf{1}$). One assertion is true: for any terminal bifurcation node (a node whose two descendants are existing species or leaves of the tree) there is an eigenvector of Σ that has elements that are positive for one of the species, negative for the other, and zero for all other species. In addition, we see that because of the block structure, every eigenvector of Σ , at a minimum, must consist of zero elements for one branch of the tree.

Beyond this, [Cavalli-Sforza and Piazza](#) describe “usual” behavior of the eigenvectors, but their ideas do not scale as the size of the tree increases. The nested block structure of Σ still has the effect of creating eigenvectors with some structure to them, though not as easily classified as suggested in [Cavalli-Sforza and Piazza \(1975\)](#). Generally the structure of the eigenvectors will not be directly related to a partition in the tree. In practice, the eigenvectors often have some relation to the bifurcations of the tree, particularly the deeper (earlier in time) bifurcations and of course the terminal bifurcations. The other eigenvectors often have clumps of positive and negative elements that correspond to subtrees of the tree, and we often empirically see as the eigenvalues get smaller some sort of concentration of large values in only a few species.