The Annals of Applied Statistics
2011, Vol. 5, No. 4, 2326–2358
DOI: 10.1214/10-AOAS402

ANALYSIS OF A DATA MATRIX AND A GRAPH: METAGENOMIC DATA AND THE PHYLOGENETIC TREE

BY ELIZABETH PURDOM[1]

University of California, Berkeley

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 nontrivial 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 illustrate that analysis of the data in a nonstandard inner-product space effectively uses this additional graphical information and produces more meaningful results.

1. Introduction.

Relationships among 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 networks. 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. 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 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 the 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.

Received November 2009; revised July 2010.

[1] Supported in part by a NSF Post-doctoral Fellowship in Biological Informatics and by NSF Grant 02-41246.

Key words and phrases. Multivariate analysis, principal components analysis, graphs, phylogenetics, metagenomics.

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 nonstandard inner-product space in $\mathbb{R}^p$ to perform the analysis of the numerical data.

The layout of the paper is as follows. First we will introduce the motivating example of bacterial composition in more detail and will return to the example at the end to demonstrate the techniques on the bacterial data. We review how PCA can be succinctly reformulated for nonstandard inner-products and its development for ecological studies of species abundance, a reformulation we will call generalized PCA (gPCA). 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 give 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. However, the same basic approach can be useful in including nonstandard forms of knowledge—other types of graphical information in particular.

Notation. In all that follows, we will use boldface type to indicate vectors and matrices and parenthetical subscripts to indicate elements of vectors and matrices. Therefore, the jth component of a vector $\mathbf{x}_i$ will be given as $\mathbf{x}_{i(j)}$ and the i, j element of a matrix $\mathbf{A}$ will be given as $\mathbf{A}_{(ij)}$.

2. Motivating example. 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 abundance.

Traditional techniques for identifying bacteria require growing the bacteria in a culture and then classifying 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 presence of different types of bacteria. The increased ease of DNA sequencing has led researchers to classify bacteria by genomic information (“metagenomics”). We focus here on the results of sequencing a specific gene (16S rDNA) found in bacteria. A random selection of all the copies of the gene present in the sample are sequenced. Ideally, each version of the gene could be uniquely identified as coming from a specific bacteria and the abundance of the different gene versions would give an estimate of the abundance of each bacteria. In reality, we do not have a direct link between a gene version and its originating bacteria, but only an estimate of it, as we explain more fully below.

Bacteria species also share an evolutionary history which might affect their biological role in the sample. We summarize the evolutionary relationship by a phylogenetic tree that describes the evolutionary history of the bacterial species. We visualize both the phylogenetic tree relating the bacterial species and their numerical abundance in Figure 1. There is a great deal of sparsity in the data; many species are present in low numbers and in only a few samples. At the same time, there are some highly abundant species 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—entire regions of the phylogenetic tree appear dissimilar between the patients, such as the *Bacteriodetes* 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 reorder the species and find other sets of species that are also very different across the patients. However, the clusters defined by the phylogenetic tree provide biological information regarding the relationships among the species that is separate from the numerical abundances. Patterns of sparsity or differences among the patients following the clusters in the tree are generally of greater interest than an arbitrary grouping since there is known biological meaning to the groupings. The additional information found by using the phylogenetic similarities can serve as a check on the kind of relationships among the species that we are interested in. This will be particularly important since we have so many more species than samples. Focusing the analysis to follow the structure of the tree will allow for more meaningful results.

This study was exploratory. It was the first sequence-based analysis of 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 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. With the small sample size, the analysis cannot extrapolate to the population in general but can only focus on describing the patients observed.

2.1. *Effect of imperfect species definition.* In practice, we cannot identify a bacterial species from the DNA sequence. Instead the sequences are themselves used to *define* the species, based on the sequence similarity of different copies—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. For this reason, the term "phylotype" is used instead of species to indicate that these are merely proxies for the true species distinctions. 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 those sequences classified into that phylotype.

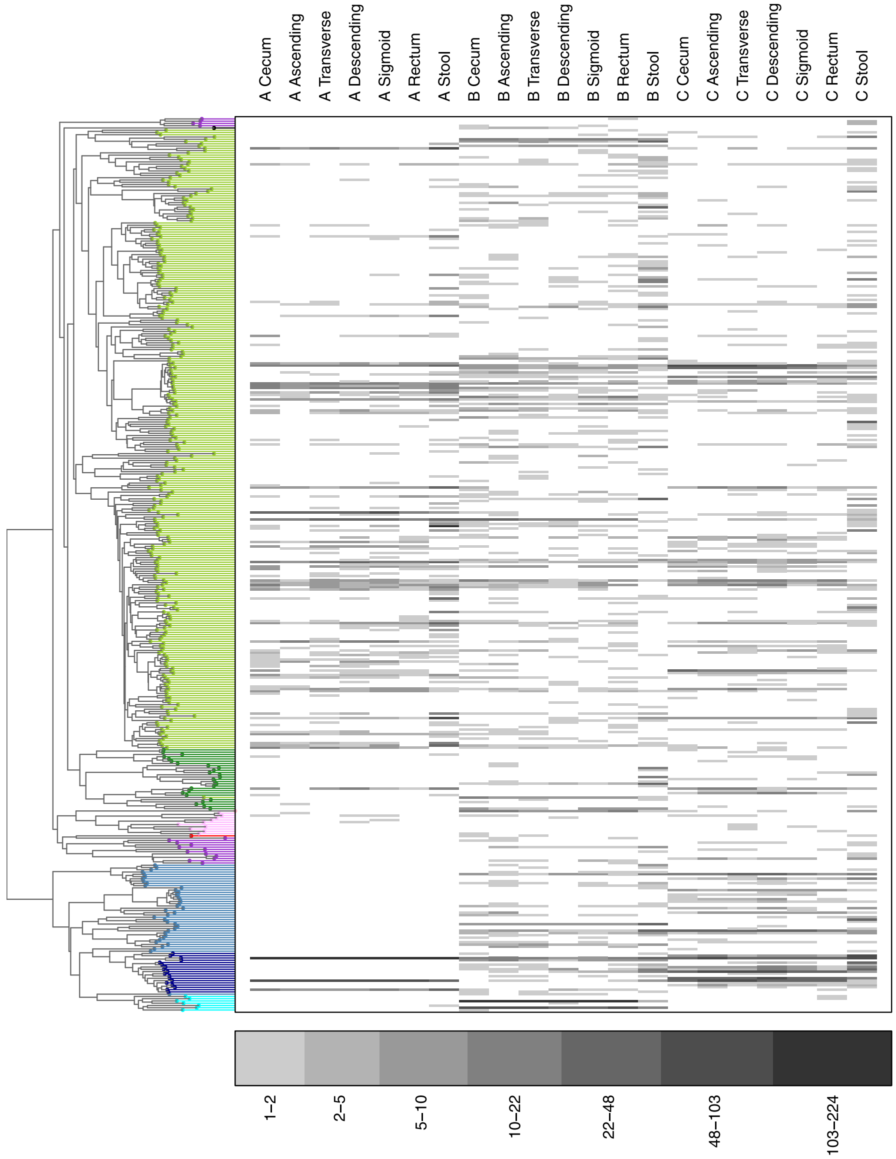


FIG. 1. *Depiction of the abundance matrix from Eckburg et al.* (2005). *Columns indicate samples, grouped by patient, and rows correspond to different phylotypes. The grey scale indicates the level of abundance on a log scale* (*see legend for conversion to original abundances*). *The colors on the phylogenetic tree indicate phylum, as in Eckburg et al.* (2005), *but with a different choice of colors: blue—Bacteriodetes, green—Firmicutes, purple—various Proteobacteria, pink—Verrucomicrobia. We additionally colored two portions of the Bacteriodetes 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*).

The possible effect of using an arbitrary cutoff for defining phylotypes is seen in Figure 1, where the length of the tree branch 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.

The relationship between DNA sequences 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 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: each of the $N = 11{,}831$ observations (or sequenced strands of DNA) belongs to one of $S = 395$ phylotypes (or species) and one of the $L = 21$ locations.

3. Incorporation of additional information via inner-products. For observed data $\mathbf{x}_i \in \mathbb{R}^p$ we propose to use nonstandard inner-products or metrics in analyzing the data. We argue that this is a simple way to include complicated outside information, such as graphical information, in the analysis of high-dimensional data.

By nonstandard inner-products, we specifically mean an inner-product between two observations i and j given by $\langle \mathbf{x}_i, \mathbf{x}_j \rangle_{\mathbf{Q}} = \mathbf{x}_i{}^T \mathbf{Q} \mathbf{y}_j$. Since $\mathbf{Q}$ also defines a metric based on $\|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{Q}}$, we may at times refer to $\mathbf{Q}$ as a metric. For any inner-product $\langle \cdot, \cdot \rangle$ and a fixed set of n vectors $\mathbf{x}_i$, there exists a matrix $\mathbf{Q}$ so that $\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \mathbf{x}_i{}^T \mathbf{Q} \mathbf{x}_j$, so this is a quite general definition. A common example of such an inner-product is the Mahalanobis distance, where $\mathbf{Q}$ is chosen as the inverse covariance matrix of the observed random vectors [see Maesschalck, Jouan-Rimbaud and Massart (2000)]. In this case, the choice of $\mathbf{Q} = \boldsymbol{\Psi}^{-1}$, where $\boldsymbol{\Psi}$ is the covariance matrix of the observed variables, removes the correlation among the variables, also known as "sphering" the data. This is the most common choice of a nontrivial $\mathbf{Q}_p$ and is used, for example, in discriminant analysis for classification problems.

The choice of an appropriate metric $\mathbf{Q}$, however, can also be a method for including outside information. In particular, assume that the additional information, such as the phylogenetic tree, is such that one can model the covariance structure $\boldsymbol{\Sigma}$ for the variables that this information would imply. The resulting covariance matrix, $\boldsymbol{\Sigma}$, is not the covariance for the observed variables in our data—which is the result of a much more complicated relationship between the graph and the data—but rather what would be expected if the data was completely created by this outside process. In order to evaluate the data so as to give priority to relationships in the phylogenetic tree, we propose using the metric $\mathbf{Q}_p = \boldsymbol{\Sigma}$ for the variable space. Performed in this space, the analysis focuses on the aspects of the data variables most congruent with the $\boldsymbol{\Sigma}$.

Because most multivariate techniques are based on inner-products, they are easily generalized to a more general inner-product space. We will focus on PCA using $\mathbf{Q}$, a technique known as generalized PCA (gPCA) or the duality principle [Escoufier (1987); Holmes (2008); Dray and Dufour (2007)]; Jolliffe (2002) gives a more in depth overview of gPCA, connecting gPCA with other techniques. We give a short review of gPCA before we discuss more fully the interpretation of this strategy. Other multivariate methods have been similarly extended and would be also relevant for incorporation of outside information.

3.1. *Generalized PCA*. Quite generally, gPCA is an ordination procedure, that is, each observed data point $\mathbf{x} \in \mathbb{R}^p$ is transformed to new, lower-dimensional data coordinates given by $\hat{\mathbf{x}} \in \mathbb{R}^k$ which is a linear transformation of the original coordinates: $\hat{\mathbf{x}} = \mathbf{Z}^T\mathbf{x}$ for some matrix $\mathbf{Z} \in \mathbb{R}^{p\times k}$. Most multivariate techniques are ordination procedures, common examples being PCA, Canonical Correlation Analysis and Correspondence Analysis. The differences lie in the choice of the linear transformation ($\mathbf{Z}$), which is chosen based on the desired properties of the new, lower-dimensional vector $\hat{\mathbf{x}}$. The most familiar example is standard PCA which seeks successive vectors $\mathbf{a}_j \in \mathbb{R}^p$ so that the resulting jth coordinate, $\hat{\mathbf{x}}_{(j)} = \langle \mathbf{x}, \mathbf{a}_j\rangle$, has the largest variance, subject to being independent of previous coordinates $\mathbf{x}_{(1)}, \dots, \mathbf{x}_{(j-1)}$; the final transformation matrix is $\mathbf{Z} = \mathbf{A}_k = (\mathbf{a}_1 \cdots \mathbf{a}_k)$.

The ordination procedure of generalized PCA (gPCA) is a generalization of PCA in that it assumes an alternative inner-product for the data vectors $\mathbf{x}$. We assume an observed random variable $\mathbf{x}$ lies in $\mathbb{R}^p$ with a known inner-product defined by $\mathbf{Q}_p \in \mathbb{R}^{p\times p}$. Then in analogy with standard principal components, gPCA can be developed from the perspective of finding the vector $\mathbf{a}$ that maximizes the population quantity, $\operatorname{var}(\langle \mathbf{a}, \mathbf{x}\rangle_{\mathbf{Q}_p})$, with $\mathbf{a}$ constrained to have unit $\mathbf{Q}_p$-norm and successive $\mathbf{a}_j$ constrained to be $\mathbf{Q}_p$-orthogonal to the preceding $\mathbf{a}_j$,

$$\|\mathbf{a}_j\|_{\mathbf{Q}_p} = 1 \quad \text{and} \quad \mathbf{A}_k^T\mathbf{Q}_p\mathbf{A}_k = \mathbf{I}_k,$$

where, again, $\mathbf{A}_k \in \mathbb{R}^{p\times k}$ is the matrix with columns $\mathbf{a}_j$. The new coordinates for $\mathbf{x}$ are then given by $\hat{\mathbf{x}} = \mathbf{A}_k^T\mathbf{Q}_p\mathbf{x}$ (so $\mathbf{Z} = \mathbf{A}_k^T\mathbf{Q}_p$ in the notation given above). As in PCA, the $\mathbf{a}_j$ will be eigenvectors, but now of the matrix $\boldsymbol{\Psi}\mathbf{Q}_p$ where $\boldsymbol{\Psi}$ is the covariance matrix of $\mathbf{x}$. The matrix $\boldsymbol{\Psi}\mathbf{Q}_p$ is not symmetric, but because $\mathbf{Q}_p$ is full rank, this is a well defined, positive definite generalized eigenequation, and the eigenvectors of $\boldsymbol{\Psi}\mathbf{Q}_p$ can be chosen to be a $\mathbf{Q}_p$-orthogonal set of vectors [see Golub and van Loan (1996)].

Just as in PCA, there are multiple developments that result in the same ordination procedure. For example, gPCA provides the best k-dimensional approximation to the inter-point similarities when the similarities are calculated in the appropriate metric space. In particular, we could note that if distances between observation i and j are given by

$$d(i, j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{Q}_p (\mathbf{x}_i - \mathbf{x}_j),$$

then gPCA is equivalent to Multidimensional Scaling (MDS) of the n observations based on these distances. Similarly, for any $\mathbf{Q}_p$, there exists a (nonunique) matrix $\mathbf{C}$ so that $\mathbf{Q}_p = \mathbf{C}\mathbf{C}^T$, which means gPCA of $\mathbf{x}$ based on $\mathbf{Q}_p$ is equivalent to first transforming the vector $\mathbf{x}$ by $\mathbf{C}$ and then performing PCA on the resulting vector $\mathbf{C}^T\mathbf{x}$.

Metric for the columns. As an analysis of a $n \times p$ data matrix $\mathbf{X}$, the above presentation only considered a metric for the space of row vectors (observations) of $\mathbf{X}$. There can also be a relevant metric for comparison of the variables, a simple example being when there are weights assigned to the observations. Generalized PCA goes beyond the description given so far and allows also for a metric $\mathbf{Q}_n \in \mathbb{R}^{n \times n}$ for the space of the column vectors of $\mathbf{X}$. These combinations of choices are generally 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 $\mathbb{R}^n$]. We note that in many cases either $\mathbf{Q}_n$ or $\mathbf{Q}_p$ are chosen to be diagonal, in which case they simplify to weights on the observations or variables, respectively.

Returning to the population development above, the inclusion of a metric for the columns of $\mathbf{X}$ is incorporated in the estimation of $\boldsymbol{\Psi}$. In order to maximize the quantity $\operatorname{var}(\langle \mathbf{a}, \mathbf{x} \rangle_{\mathbf{Q}_p})$, we must estimate $\boldsymbol{\Psi}$ from our data matrix $\mathbf{X}$; we include the metric $\mathbf{Q}_n$ for the columns in our estimate so that $\widehat{\boldsymbol{\Psi}} = \mathbf{X}^T \mathbf{Q}_n \mathbf{X}$. Then our estimates of $\mathbf{a}_j$ are given by the eigenvectors of $\mathbf{X}^T \mathbf{Q}_n \mathbf{X} \mathbf{Q}_p$. A geometric development that includes the metric $\mathbf{Q}_n$ for the columns shows that gPCA best preserves the total inner-point similarities of the data matrix $\mathbf{X}$ when using a measure of inner-point similarities incorporating the row and column matrix known as the inertia (see Appendix C). In addition to the geometric view of gPCA, Jolliffe (2002) notes that gPCA with $\mathbf{Q}_n$ a diagonal matrix provides the maximum likelihood estimates of the fixed effects version of a factor model,

$$\mathbf{x} = \mathbf{A}\mathbf{z} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{Q}_n^{-1} \mathbf{Q}_p^{-1})$.

Connection between analysis of the rows and columns. In some data settings either the rows or the columns can be meaningfully considered as the observations, such as analysis of large contingency tables that are our motivating example. Furthermore, the importance of the different variables in describing a low-dimensional representation of the observations is a common part of PCA. A gPCA of the *columns* of $\mathbf{X}$, also reduced to k dimensions, is technically the gPCA of triplet $(\mathbf{Y} = \mathbf{X}^T, \mathbf{Q}_n, \mathbf{Q}_p)$ and results in new coordinates for the columns given by $\hat{\mathbf{Y}} = \mathbf{Y}\mathbf{Q}_n\mathbf{B}_k \in \mathbb{R}^{p \times k}$.

Again in analogy to PCA, a generalized form of the SVD of $\mathbf{X}$ yields the solutions to gPCA on both the columns or the rows simultaneously. If the rank of $\mathbf{X} = r$, we can write $\mathbf{X} = \mathbf{B}\boldsymbol{\Lambda}^{1/2}\mathbf{A}^T$, where $\mathbf{A} \in \mathbb{R}^{p\times r}$ and $\mathbf{B} \in \mathbb{R}^{n\times r}$, and 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 eigenequations are

$$\mathbf{X}^T\mathbf{Q}_n\mathbf{X}\mathbf{Q}_p\mathbf{A} = \mathbf{A}\boldsymbol{\Lambda},$$

$$\mathbf{X}\mathbf{Q}_p\mathbf{X}^T\mathbf{Q}_n\mathbf{B} = \mathbf{B}\boldsymbol{\Lambda},$$

and for any choice of k, $\mathbf{B}_k = \mathbf{X}\mathbf{Q}_p\mathbf{A}_k\boldsymbol{\Lambda}_k^{-1/2}$, where $(\cdot)_k$ refers to the matrix with the first k columns or diagonal elements, as appropriate.

This means the new coordinates from a gPCA of the rows can be completely determined by the new coordinates from a gPCA of the columns of the data matrix. Let $\hat{\mathbf{x}} \in \mathbb{R}^k$ be the new coordinates for a vector $\mathbf{x} \in \mathbb{R}^p$ based on the gPCA of the rows of $\mathbf{X}$. The new coordinates are given as

$$\hat{\mathbf{x}}^T = \mathbf{x}^T\mathbf{Q}_p\hat{\mathbf{Y}}\boldsymbol{\Lambda}_k^{-1/2}.$$

Put another way, the value of the j new coordinates of $\hat{\mathbf{x}}$ is given by

$$\hat{\mathbf{x}}_{(j)} = \langle \mathbf{x}, \chi_j\rangle_{\mathbf{Q}_p},$$

where χ_j is the jth column of $\hat{\mathbf{Y}}\boldsymbol{\Lambda}_k^{-1/2}$, that is, the column of $\hat{\mathbf{Y}}$ normalized to have standard deviation one. Thus, the jth coordinate of $\hat{\mathbf{x}}$ is a measure of the similarity of $\mathbf{x}$ with the jth variable defining the reduced space of the columns.

3.2. *Interpretation of nonstandard metrics.* Using a metric for $\mathbb{R}^p$ has an obvious rationale when the metric is a diagonal, implying different weights for different variables, or when the metric is $\boldsymbol{\Psi}^{-1}$ where $\boldsymbol{\Psi}$ is the covariance of the variables (Mahalanobis distance). However, it is not immediately clear why a particular matrix $\mathbf{Q}_p$, such as $\mathbf{Q}_p = \boldsymbol{\Sigma}$ as we propose above, would improve a given data analysis. 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 the perspective of Rapaport et al. (2007) in their proposal for the particular case of general graphs (see Section 7).

Outside information, such as our phylogeny, when represented by $\boldsymbol{\Sigma}$ also defines a basis given by the eigenvectors $\mathbf{v}_j$ of $\boldsymbol{\Sigma}$. The eigenvectors decompose our overall covariance into hopefully informative directions with regards to our outside structure, and the $\mathbf{v}_j$ can be ordered based on their overall contribution to $\boldsymbol{\Sigma}$ based on the eigenvalues λ_j. The directions given by the $\mathbf{v}_j$ can be weighted in different ways to create a family of metrics, with each choice of weighting system emphasizing different directions.

More precisely, suppose $\mathbf{\Sigma}$ has an eigendecomposition given by $\mathbf{V\Lambda V}^T$; $\mathbf{V}$ is a $p \times p$ matrix with columns $\mathbf{v}_j$ consisting of the eigenvectors of $\mathbf{\Sigma}$, and $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues λ_j. The vectors $\mathbf{v}_j$ form a basis for $\mathbb{R}^p$ and, therefore, a data vector $\mathbf{x}$ can be written as

$$\mathbf{x} = \sum_j \langle \mathbf{v}_j, \mathbf{x} \rangle \mathbf{v}_j = \mathbf{V}\breve{\mathbf{x}},$$

where $\breve{\mathbf{x}}_{(j)} = \mathbf{v}_j{}^T\mathbf{x}$ gives the magnitude of $\mathbf{x}$ in the direction of the eigenvectors of $\mathbf{\Sigma}$.

This decomposition of $\mathbf{x}$ into its contributions due to the directions given by $\mathbf{v}_j$ creates no loss of information, being only a change of basis. But we can transform the original $\mathbf{x}$ by giving weights $\mathbf{w}_{(j)}$ to different directions in order to give more emphasis to the features that $\mathbf{v}_j$ represents, in which case we now have a new vector $\mathbf{f_w} \in \mathbb{R}^p$ with

$$\mathbf{f_w}(\mathbf{x}) = \sum_j \mathbf{w}_{(j)}\breve{\mathbf{x}}_{(j)}\mathbf{v}_j = \mathbf{VD_w}\breve{\mathbf{x}},$$

where $\mathbf{D_w}$ is the diagonal matrix with diagonal given by $\mathbf{w}$. For example, if our outside structure could be represented in a smaller subspace so that $\mathbf{\Sigma}$ had rank $r < p$, then defining $\mathbf{w}_{(j)} = \mathbb{1}\{j \le r\}$ would give $\mathbf{f_w}(\mathbf{x})$ as the projection of $\mathbf{x}$ onto the smaller subspace defined as relevant by our outside structure. More generally, the eigenvalues λ_j quantify the contribution of a direction $\mathbf{v}_j$ to our outside structure $\mathbf{\Sigma}$, and, therefore, the eigenvalues, or a monotone transformation of them, are a smoother way to assign relative importance to the different basis defined by $\mathbf{\Sigma}$.

For two vectors $\mathbf{x}$ and $\mathbf{y}$, the standard inner-product between $\mathbf{f_w}(\mathbf{x})$ and $\mathbf{f_w}(\mathbf{y})$ is given by

$$\langle \mathbf{f_w}(\mathbf{x}), \mathbf{f_w}(\mathbf{y}) \rangle = \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{VD}_\mathbf{w}^2\mathbf{V}^T},$$

that is, the inner-product between $\mathbf{x}$ and $\mathbf{y}$ using the metric $\mathbf{VD}_\mathbf{w}^2\mathbf{V}^T$. Then the choice of a metric $\mathbf{Q}_p = \mathbf{\Sigma}$ is equivalent to the choice of weighting each $\mathbf{v}_j$ by $\lambda_j^{1/2}$ and $\mathbf{f_w}(\mathbf{x}) = \mathbf{Q}_p^{1/2}\mathbf{x}$.

In this light, we can compare the effect of using $\mathbf{Q}_p = \mathbf{\Sigma}$ versus $\mathbf{Q}_p = \mathbf{\Sigma}^{-1}$. Both obviously have the same eigenvectors and differ only in the weighting the eigenvectors (λ_j versus $1/\lambda_j$). Thus, the choice of $\mathbf{\Sigma}$ as the metric for the variables places emphasis on the directions with more information about the outside structure, while $\mathbf{\Sigma}^{-1}$ emphases directions that are most independent of the outside information. Depending on whether this outside structure is thought to enlighten or confound the analysis, the different weighting systems are appropriate.

From this harmonic perspective, the behavior of the eigenvectors is quite revealing as to the intuitive interpretation that can be placed on the analysis. Such a projection onto a relevant set of basis is, of course, analogous to harmonic analysis or wavelet analysis for functional data. PCA could also be described similarly, only with the $\mathbf{v}_j$ 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 for the analysis in question is captured in the first few directions. More generally, eigenvectors of a covariance matrix describe linear combinations of decreasing variance, and thus presumably decreasing ability to reveal the structure of interest.

Beyond the ordering of the eigenvectors, a desirable behavior for the purposes of interpretability is for the bases (eigenvectors) to be sparse—nonzero in a small portion of the coordinate space (or, more generally, a clearly interpretable subspace). If so, the resulting coordinates of the transformed data are easily interpreted as contrasts or combinations of a small set of variables. This is the appeal of wavelets or various sparse PCA algorithms. From the point of view of our outside information in the form of a graph or phylogenetic tree, this means we want our representation of the outside information (via $\mathbf{\Sigma}$) to result in eigenvectors that are interpretable decompositions of the external information we have. As we will see, certain covariance structures for phylogenies and also graphs have such decompositions, which is one reason that the analysis in a nonstandard inner-product space can give highly interpretable results.

3.3. *gPCA and analyses of variables as observations.* Another interpretation of $\mathbf{Q}_p$ slightly different from the geometric one given above is that it is simply an additional data matrix—one that defines similarities between the p variables—which we wish to include into our analysis of the primary data matrix, $\mathbf{X}$.

Pavoine, Dufour and Chessel (2004) accomplish this by their method of Double Principal Coordinates Analysis (DPCoA), which explicitly transforms the similarities between the variables given by $\mathbf{Q}_p$ into a set of standard Euclidean coordinates, $\mathbf{Z} \in \mathbb{R}^{p \times r}$, using MDS (also known as Principal Coordinates Analysis). This can be viewed as giving an alternative basis for $\mathbb{R}^p$ and $\mathbf{Z}$ as the new set of coordinates of the original p variables in which $\mathbf{X}$ was measured. Then the next step of DPCoA transforms the data $\mathbf{X}$ to this new basis as well, that is, to coordinates $\mathbf{XZ}$. DPCoA then performs PCA on the transformed $\mathbf{X}$ (we note that these steps are exactly the same as the steps of DPCoA, but generalized here to apply to general data matrices $\mathbf{X}$ and not just the contingency tables originally proposed; see Appendix A for details).

The series of steps that make up DPCoA is exactly equivalent to a single gPCA of the centered data matrix, $\tilde{\mathbf{X}}$, with the choice of metrics given by the triplet $(\tilde{\mathbf{X}}, \mathbf{Q}_p, \mathbf{Q}_n)$, provided that (1) the centered data matrix of $\mathbf{X}$ was the result of centering the *columns* (variables) and (2) the same centering matrix used in centering $\mathbf{X}$ was also used in the MDS of $\mathbf{Q}_p$ to find the matrix $\mathbf{Z}$ (Appendix A). DPCoA was only proposed for the particular setting of ecological studies where the data is a contingency table, and, thus, centering the columns of $\mathbf{X}$ is actually equivalent to centering the rows because of the row and column weights that are typically chosen for the centering (see Section 4.2), so the requirement is naturally satisfied.

By recasting DPCoA as a gPCA, the technique now has general application and is clearly extendable, since in many situations heterogenous information can be similarly introduced into an analysis in this way.

We note that MDS is traditionally described based on an input of squared dissimilarities or distances between points given by a $p \times p$ matrix $\boldsymbol{\delta}$; however, any positive definite $\mathbf{Q}_p$ that can be written as

$$\mathbf{Q}_p = \mathbf{1}_p \mathbf{v}^T + \mathbf{v}\mathbf{1}_p^T - \tfrac{1}{2}\boldsymbol{\delta}$$

for some vector $\mathbf{v} \in \mathbb{R}^p$ will result in the same MDS of the variables and thus the same DPCoA results.

Another approach to analyzing two sources of data are multivariate kernel techniques, such as kernel CCA [Bach and Jordan (2002)], which assume that the only knowledge of the data is similarities between objects. In these techniques, two sets of data provide two different sets of kernel similarity matrices $\mathbf{K}_1$ and $\mathbf{K}_2$ on the same set of n objects, and the kernel analysis results in new coordinates $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$ that are linear combinations of these kernel similarities that best relate the two data sets (the prediction context is also possible). Then gPCA of the rows of $\mathbf{X}$ results in equivalent coordinates for the rows as the choice of $\mathbf{K}_1 = \mathbf{X}\mathbf{Q}_p\mathbf{X}^T$ and $\mathbf{K}_2 = \mathbf{Q}_n$, for an extreme form of regularization of the CCA problem that only constrains the norm $\|f\|^2$ of the resulting functional, rather than the more common constraint on estimated variance (see Appendix B).

In the current setting, we are instead interested in outside information on the p variables in the form of $\mathbf{Q}_p$. In this case, the natural kernel analysis would provide new coordinates for the p columns based on $\mathbf{K}_1 = \mathbf{X}^T\mathbf{Q}_n\mathbf{X}$ and $\mathbf{K}_2 = \mathbf{Q}_p$, which would correspond to a gPCA of the columns. As we noted above, however, the row coordinates from a gPCA of the rows are recoverable from the gPCA of the columns. Like DPCoA, this perspective of gPCA is that of finding a new set of coordinates for the variables, based this time on explicitly relating the expected similarities to the observed similarities, and then rotating the matrix $\mathbf{X}$ into this basis.

4. Analysis of species abundance.

The investigation of species composition and comparison of species across different locations, such as in our motivating example of the bacteria communities, 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. Development of gPCA as described here has often been in this setting, thus it is useful to review some important points before returning to our bacteria example.

Our motivating example of the bacteria is 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.* Assume that the abundance of certain species are measured at L different locations and a total of S distinct species types are observed. We drop the use of n and p for the rows and columns of our data matrix to emphasize that there is not a canonical dimension that is considered the observations in this setting, though we will focus on the locations as observations in our example. We will similarly use matrices $\mathbf{Q}_S$ and $\mathbf{Q}_L$ for the row and column metrics.

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 $\mathbf{x}_\ell$ 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 $\mathbf{X} \in \mathbb{R}^{L\times S}$. Namely, let $\mathbf{w}_L = \mathbf{A1}/N \in \mathbb{R}^L$ be the row sums of $\mathbf{A}$ normalized to sum to one. Then $\mathbf{X}$ is given by

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

where $\mathbf{D}_{\mathbf{w}_L}$ is a diagonal matrix with diagonal elements given by $\mathbf{w}_L$ respectively.

The vector $\mathbf{w}_L$ also defines weights for each of the locations, and the weights are proportional to the total number of observations in that location. The weighted mean of the locations, $\bar{\mathbf{x}}$, is given by $\mathbf{X}^T\mathbf{w}_L$ and the centered data matrix, $\widetilde{\mathbf{X}}$, is given by $\widetilde{\mathbf{X}} = (\mathbf{I} - \mathbf{1}\mathbf{w}_L^T)\mathbf{X}$.

4.2. *A few important properties of contingency tables. The duality of rows and columns.* Note that the weighted mean, $\bar{\mathbf{x}}$, also sums to one and therefore is itself a potential location profile. In fact, $\bar{\mathbf{x}}$ is proportional to the column sums of $\mathbf{A}$ and thus is equal to the relative frequency of the species across *all* locations. If we had instead chosen to analyze the columns (species) as the observations, choosing weights $\mathbf{w}_S$ for the species in the same way as the rows, we would have $\mathbf{w}_S = \bar{\mathbf{x}}$.

The equivalence of $\mathbf{w}_S$ and $\bar{\mathbf{x}}$ has interesting repercussions for data analysis because under these weighting schemes, we can equivalently center either the rows or the columns,

$$\widetilde{\mathbf{X}} = \mathbf{P}_{\mathbf{w}_L}\mathbf{X} = \mathbf{X}\mathbf{P}_{\mathbf{w}_S},$$

where $\mathbf{P}_{\mathbf{w}_m} = (\mathbf{I}_m - \mathbf{1}_m\mathbf{w}_m^T)$ is the projection matrix that centers m observations based on a weighted mean with $\mathbf{w}_m$ as weights.

Interpretation of variables in gPCA. Because we analyze location profiles, there is a simple 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, not just as a direction in space, so we can legitimately visualize distances between the location and species in a single plot.

Examples of gPCA with contingency tables. In addition to DPCoA described above, different metric spaces are often used for analyzing contingency tables via gPCA, particularly to retain additional information such as the weights $\mathbf{w}_L$ and/or $\mathbf{w}_S$. 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 $(\widetilde{\mathbf{X}}, \mathbf{D}_{\mathbf{w}_S}^{-1}, \mathbf{D}_{\mathbf{w}_L})$ [see Greenacre (1984) for a detailed treatment]. This gives an inner product of the form $\mathbf{x}_k^T \mathbf{D}_{\mathbf{w}_S}^{-1} \mathbf{x}_\ell$, 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.

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 are probably less trustworthy. Gimaret-Carpentier, Chessel and Pascal (1998) propose no weighting of the species, only the locations, which gives a triplet $(\widetilde{\mathbf{X}}, \mathbf{I}_S, \mathbf{D}_{\mathbf{w}_L})$—just a regular PCA with weights on each observation. Such an analysis in ecology is also called Non-symmetric Correspondence Analysis (NSCA).

4.3. *Connection to diversity.* We take a moment to comment on the connection of the choice of gPCA metrics to a common question in ecology—how "diverse" a location is. Diversity is a measurement of how close the distribution of species is to uniform. Two popular measures of diversity are variations of the Gini–Simpson index, $H_{\mathrm{GS}}(\mathbf{x}) = 1 - \sum_{s=1}^{S} \mathbf{x}_{(s)}{}^2$, and the Shannon Diversity index, $H_{\mathrm{Sh}}(\mathbf{x}) = \sum_{s=1}^{S} \mathbf{x}_{(s)} \log(\mathbf{x}_{(s)})$.

Ecology studies often use the individual diversity of locations to make comparisons, but the diversity indices alone do not effectively compare the species composition. Locations can have quite different composition of species but with same levels of individual diversity. Of interest is how the species composition changes, and ordination techniques are used to address these problems, but as a separate component of the analysis of the ecological data. However, the choice of diversity and the choice of gPCA parameters are closely connected, as pointed out in Pélissier et al. (2003). Namely, if $\mathbf{Q}_L$ is a simple diagonal matrix of weights on the locations, gPCA of $(\mathbf{X}, \mathbf{Q}_S, \mathbf{Q}_L)$ gives the best representation of a particular measure of dissimilarity between locations, and choice of this dissimilarity measure implies a diversity measure, and vice versa. Pélissier et al. (2003) stated this for several specific ordination techniques, and we state it more generally for any choice of metric $\mathbf{Q}_S$ on $\mathbb{R}^S$. Define diversity and dissimilarity measures for any positive definite matrix $\mathbf{Q}_S = \mathbf{Q}$ by

$$H_{\mathbf{Q}}(\mathbf{x}) = \mathbf{x}^T \operatorname{diag}(\mathbf{Q}) - \mathbf{x}^T \mathbf{Q}\mathbf{x} = \sum_r \mathbf{x}_{(r)} \mathbf{Q}_{(rr)} - \sum_{rs} \mathbf{Q}_{(rs)} \mathbf{x}_{(r)} \mathbf{x}_{(s)},$$

$$\operatorname{Diss}_{\mathbf{Q}}(\mathbf{x}_k, \mathbf{x}_j) = (\mathbf{x}_k - \mathbf{x}_\ell)^T \mathbf{Q} (\mathbf{x}_k - \mathbf{x}_\ell).$$

These are clearly closely related to the norm and inner-product defined with the choice of $\mathbf{Q}$. With these choices of diversity and dissimilarity, the total diversity across all locations is given by $H_{\mathbf{Q}}(\bar{\mathbf{x}})$ and can be decomposed into the average diversity of individual locations and plus the average of pairwise dissimilarities of locations,

$$\underbrace{H_{\mathbf{Q}}(\bar{\mathbf{x}})}_{I_{\text{Total}}} = \underbrace{1/2 \sum_{k=1}^{L} \sum_{\ell=1}^{L} \mathbf{w}_{L(k)} \mathbf{w}_{L(\ell)} \operatorname{Diss}_{\mathbf{Q}}(\mathbf{x}_\ell, \mathbf{x}_\ell)}_{I_{\text{Between}}} + \underbrace{\sum_{\ell=1}^{L} \mathbf{w}_{L(\ell)} H_{\mathbf{Q}}(\mathbf{x}_\ell)}_{I_{\text{Within}}}.$$

gPCA of $(\tilde{\mathbf{X}}, \mathbf{Q}, \mathbf{D}_{\mathbf{w}_L})$ gives the best low-dimensional representation of I_{B}, the average dissimilarity between locations (see Appendix C).

We can define a F-style statistic, as in ANOVA, to test for significant dissimilarity between the locations [Legendre and Legendre (1998)]

$$F = \frac{(N-1) I_{\mathrm{B}}}{L I_{\mathrm{W}}}.$$

Because the significance of F will generally be determined by permutation tests, this F-test is functionally equivalent to using $I_{\mathrm{B}}/I_{\mathrm{T}}$, which has many appealing connections to standard measures. We describe a few of them below given originally by Pélissier et al. (2003) and Pavoine, Dufour and Chessel (2004):

CA: For correspondence analysis, $\mathbf{Q} = \mathbf{D}_{\mathbf{w}_S}^{-1}$ results in a dissimilarity between profiles measured by the χ^2 distance,

$$(\mathbf{x}_k - \mathbf{x}_\ell)^T \mathbf{D}_{\mathbf{w}_S}^{-1} (\mathbf{x}_k - \mathbf{x}_\ell),$$

which has also been proposed for document classification. As is well known in CA, $I_{\mathrm{B}} = \chi^2/N$, where χ^2 is the χ^2-statistic for testing independence. The implied diversity measurement for a profile $\mathbf{x}$ is $\sum \mathbf{w}_{S(r)} \mathbf{x}_{(r)} (1 - \mathbf{x}_{(r)})$, which implies the total diversity I_{T} is simply $S - 1$. Thus, $I_{\mathrm{B}}/I_{\mathrm{T}}$ is proportional to the χ^2 statistic.

DPCoA: As we saw before, DPCoA can be written in terms of a general $\mathbf{Q}_S$. If we write $\mathbf{Q}_S = \mathbf{1}_p \mathbf{v}^T + \mathbf{v} \mathbf{1}_p^T - \frac{1}{2} \boldsymbol{\delta}$ for some $\mathbf{v} \in \mathbb{R}^S$ and species dissimilarities $\boldsymbol{\delta}$, as in Section 3.3, then we have that $H_{\mathbf{Q}}$ and $\operatorname{Diss}_{\mathbf{Q}}$ are the Rao diversity and dissimilarity measures [Rao (1982)] given by

$$H_{\mathbf{Q}}(\mathbf{x}) = \sum_{rs} \boldsymbol{\delta}_{(rs)} \mathbf{x}_{(r)} \mathbf{x}_{(s)},$$

$$\mathrm{Diss}_{\mathbf{Q}}(\mathbf{x}_k, \mathbf{x}_j) = (\mathbf{x}_k - \mathbf{x}_\ell)^T \left(-\tfrac{1}{2}\boldsymbol{\delta}\right)(\mathbf{x}_k - \mathbf{x}_\ell).$$

Thus, gPCA with $\mathbf{Q}_S$ results in differences between locations profiles being down-weighted for the species that are similar to each other and up-weighted for very distinct species. Though stated in many individual steps and not a single gPCA as we do here, the DPCoA method was motivated by searching for an ordination that maximized this notion of distance between observations. The ratio $I_{\mathrm{B}}/I_{\mathrm{T}}$ is commonly called the F_{ST} statistic [Martin (2002)] in biological applications and has been suggested for testing differences in bacterial communities, where $\boldsymbol{\delta}$ is usually chosen as the original measures of genetic distance between the sequences. The F_{ST} statistic is also used in testing for differences of allele composition in human populations [Excoffier, Smouse and Quattro (1992)].

NSCA: Since NSCA is standard PCA, except for the weighting of the observations, $\mathbf{Q} = \mathbf{I}$, and is equivalent to the Rao diversity and dissimilarity measures when all the species are equally distant from each other. The resulting measure of diversity in this case is the Gini–Simpson measure of diversity, H_{GS}. The ratio $I_{\mathrm{B}}/I_{\mathrm{T}}$ is equivalent to Kendall's τ [D'Ambra and Lauro (1992)].

5. A metric for species related by a phylogenetic tree. Returning to our bacteria example, we want a matrix $\boldsymbol{\Sigma}$ that represents the phylogenetic relationships of the species. As mentioned in Section 3, if we can model the covariance structure of data expected based on just our outside information, this provides a natural choice of $\boldsymbol{\Sigma}$. The phylogenetic tree in fact is a representation of the process of evolution, for which many possible probabilistic models could be created.

A common probabilistic model for the evolution of the value of a trait over time, due to Cavalli-Sforza and Piazza (1975), is one of a Brownian motion model over time, where at each speciation event the model assumes that the resulting sister species continue to evolve independently [for alternative models of evolution, see Hansen and Martins (1996); Pavoine et al. (2008)]. This model gives a covariance structure for the trait as observed on the existing species (the leaves of the phylogenetic tree) and can be simply stated in terms of distances between species on the phylogenetic tree. Moreover, the eigenvectors of this covariance matrix generally demonstrate nice localization properties relative to the tree, implying interpretable results in terms of the properties of the tree.

Specifically, assume that there is a known phylogenetic tree describing the ancestral relationship of S extant species and that a trait of interest for these species has evolved over time according to the model of independent Brownian motion with the speciation as depicted on this tree. The S extant species are observed, and for each species s at a single time point $\mathbf{t}_{(s)}$, the trait is measured, resulting in $\mathbf{y}_{(s)}$.

Then the vector of trait values, $\mathbf{y}$, follows a multivariate normal distribution with covariance between species r and s proportional to the total length of time that the evolutionary history of the two species were identical, $\mathrm{cov}(\mathbf{y}_{(s)}, \mathbf{y}_{(r)}) = \sigma^2 t_{rs}$, where t_{rs} is the time at which the two lineages diverged, as measured from their most common ancestor.

We 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 $\boldsymbol{\delta}$ be the distance matrix of the leaves based on the distance of the shortest path between them on the tree. Then we can write the covariance matrix $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma} = 1/2(\mathbf{1}\mathbf{t}^T + \mathbf{t}\mathbf{1}^T - \boldsymbol{\delta}),$$

where $\mathbf{t} \in \mathbb{R}^S$ is the vector of the distance of each species to the root.

This relationship between $\boldsymbol{\Sigma}$ and $\boldsymbol{\delta}$ implies that gPCA with $\boldsymbol{\Sigma}$ as the species metric will decompose a Rao Dissimilarity, with dissimilarities between species given as their distance on the tree. For the bacterial example, use of this distance has the effect of not declaring locations very different if the differences between locations occur in phylogenetically similar phylotypes.

Properties of phylogenetic metric. We would like that the eigenvectors of $\boldsymbol{\Sigma}$ 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 multiscale 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 $\boldsymbol{\Sigma}$ have this multiscale structure [e.g., Cavalli-Sforza and Piazza (1975); Rohlf (2001); Martins and Housworth (2002)], but 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 have some characteristics of this multiscale property; for example, $\boldsymbol{\Sigma}$ has a block structure which guarantees that the eigenvectors of $\boldsymbol{\Sigma}$ will, at a minimum, be nonzero for only one side or the other of the initial split in the tree (Appendix E). 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 closely-related groups based on the sign of the entries, though the groups do not exactly correspond to subtrees (see Figure 2).

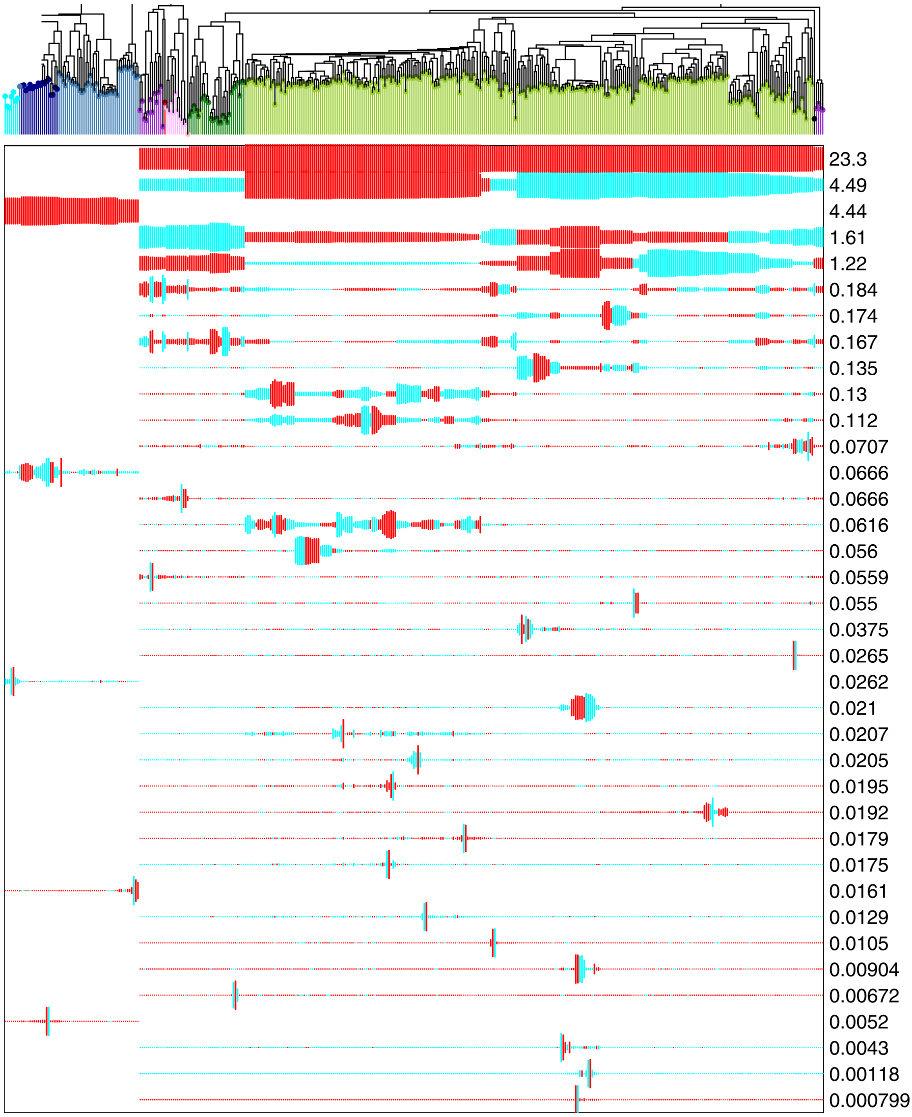
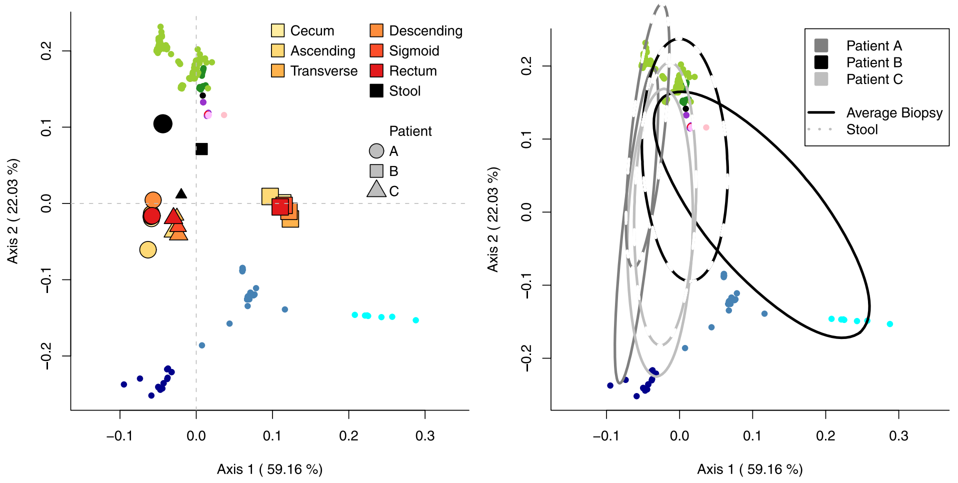


FIG. 2. *An illustration of some eigenvectors of* $\mathbf{\Sigma}$ *for the intestinal data. Only* 25 *of the* 395 *eigenvectors are shown: those that correspond to the first five largest eigenvalues, the last two smallest eigenvalues, and then a random sample in between. Each row represents an eigenvector, and the value of each element of the vector is plotted alongside the phylotype with which it corresponds. Blue represents a positive value, red a negative. The width indicates the absolute value of the element. Again, each row has been normalized so that the maximum width is the same in each row. Next to each row is printed the corresponding eigenvalue.*



(a) (b)

FIG. 3. *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.*

6. gPCA applied to bacterial data and phylogenetic tree. In Eckburg et al. (2005) our original analysis of the bacterial data was a gPCA of $(\tilde{\mathbf{X}}, \boldsymbol{\Sigma}, \mathbf{D}_{\mathbf{w}_L})$, which is equivalent to DPCoA choosing $\boldsymbol{\delta}$ to be the distance among the phylotypes. We display in Figure 3 the ordination of the locations (samples) and species using the first two coordinates (using the implementation of DPCoA 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 their value when projected onto 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 to other common choices of metrics 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 3 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.

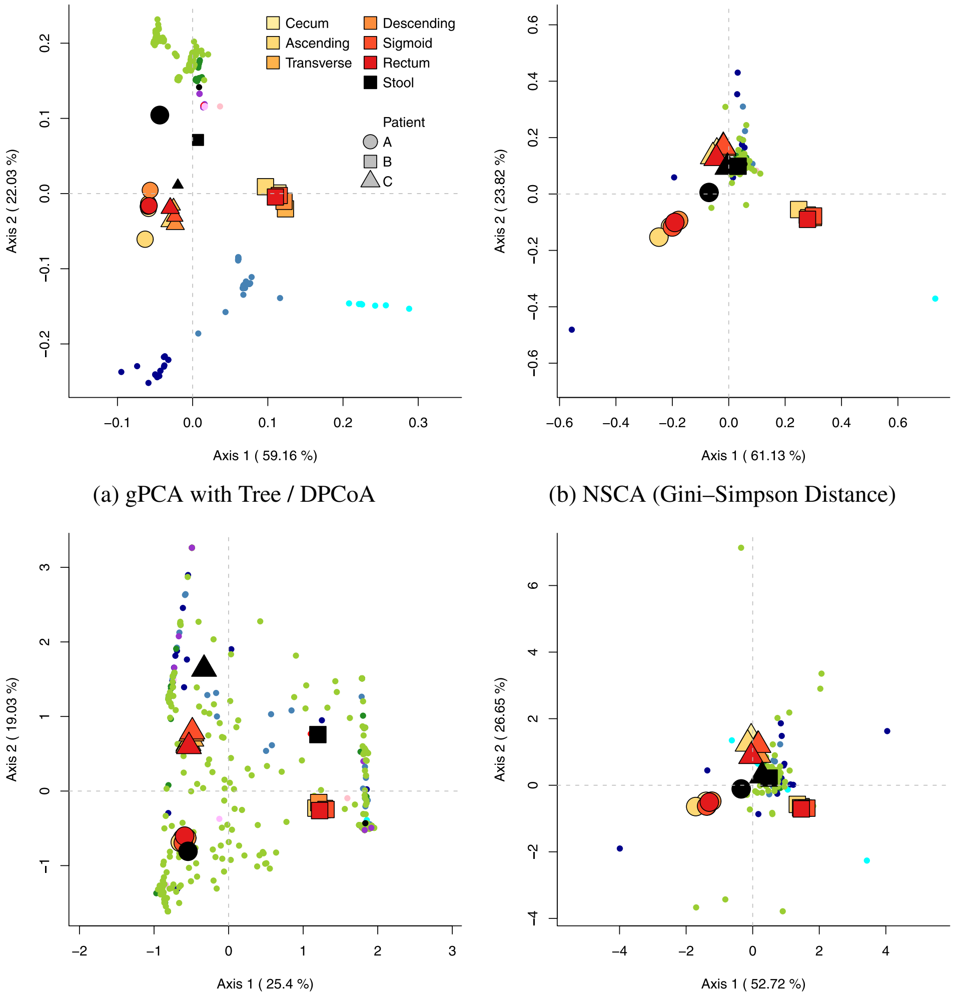
The most striking aspect of the plot from the gPCA is the additional information provided from the inclusion of the phylotypes in the plot. Recall that when our data matrix $\mathbf{X}$ consists of profile vectors, our original axes $\mathbf{e}_s$ correspond to a location that is entirely concentrated in phylotype s. The coordinates of the phylotypes given by gPCA will be the coordinates of our axis $\mathbf{e}_s$ centered and rotated like the observed profiles (see Appendix A). Looking at the ordination 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 3(b) depicts the different samples as ellipses with the axes of the ellipses determined by the relative proportion of the different species for the location (see Appendix F). 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 (e.g., 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 gPCA. These dimensions do cover a large proportion of the Rao Dissimilarity, but still are only an approximation of the full space. We are mainly focused on demonstrating the characteristics of the ordination procedure in terms of the coordinate system that it creates, but for more rigorous testing of differences between the patients or between the biopsies and stool samples, permutation tests based on the F-statistic described above would generally want to compare with the entire coordinate system.

6.1. *Comparison to other approaches*. How do these results compare to the other ordination techniques mentioned above? In Figure 4 we show the results of the ordination from Non-symmetric Correspondence Analysis (NSCA), Correspondence Analysis (CA) and a Mahalanobis-like distance based on $\mathbf{\Sigma}^{-1}$ (see Section 5). We similarly center, rotate and project the axes $\mathbf{e}_s$ 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 the gPCA using the tree 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 our gPCA but with each species at equal



(a) gPCA with Tree / DPCoA

(b) NSCA (Gini–Simpson Distance)

(c) CA

(d) gPCA with $\mathbf{\Sigma}^{-1}$

FIG. 4. *Coordinates of species and samples from alternative ordination techniques.*

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 [Figure 4(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 among 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. This approach focuses the analysis on finding an interpretation among 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 problem 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.

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

In Figure 5 we order the elements of $\mathbf{v}_i$ 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 $\mathbf{\Sigma}$ 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

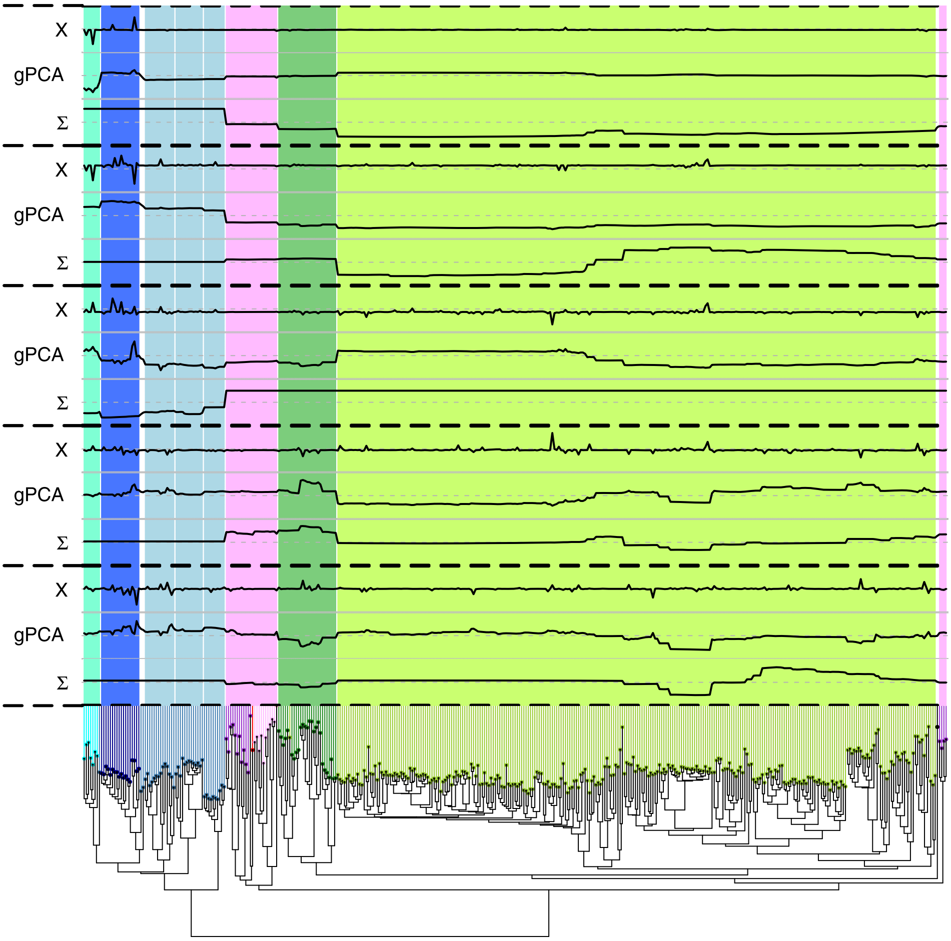


FIG. 5. *Shown are the first five linear combinations of gPCA using* $\mathbf{\Sigma}$ *that act on the observations in* $\mathbf{X}$ (*the location profiles*) *to create the first five coordinates* ($\mathbf{v}_i$). *The five dimensions are divided by thick, dotted line. Also shown adjacent to each gPCA vector are the linear combinations from a standard PCA of* $\widetilde{\mathbf{X}}$ (*labeled 'X'*) *and the eigenvectors of* $\mathbf{\Sigma}$ (*labeled '*$\mathbf{\Sigma}$*'*).

the phylotypes as variables and there are far too many variables for the number of samples we have.

Despite the intuitive results, the analysis depends on our choice of encoding the tree using $\mathbf{\Sigma}$ (or, equivalently, for DPCoA, our choice of $\boldsymbol{\delta}$). In particular, the block structure of $\mathbf{\Sigma}$ 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 $\boldsymbol{\delta}$, where these two lineages will be far away from each other and, thus, differences between will be accorded more weight in the analysis. However, as mentioned above, we see that the method depends only on $\boldsymbol{\delta}$, so the definition of the root of the tree, per se, is not the deciding factor, but rather the large amount of distance between these two subtrees.

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.

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, where the the laplacian matrix $\mathbf{L}$ is given by $\mathbf{D}_{\mathbf{d}} - \mathbf{A}$, where $\mathbf{A}$ is the adjacency matrix of the graph and $\mathbf{d}$ is the vector of degrees of each node. The Laplacian matrix is a natural choice for graphs; the eigenvectors have similar multiscale properties as our metric for the phylogenetic tree. In Appendix D 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.

A related application is found in spatial analysis, where spatial relationships between observations are based on neighborhood relationships, or, more generally, distances between points. While many analyses first remove the spatial dependencies so as to have independent observations, it is often also of interest to evaluate the relationship between the spatial patterns and the observed data. When spatial connectivity between observations is simplified to a zero–one connectivity measure (usually based on a cutoff on the distance between the observations), the spatial relationship is given by an adjacency matrix. Geary's c and Moran's I, two common measures of the spatial autocorrelation of $\mathbf{y} \in \mathbb{R}^n$ (a variable observed on the n observations), can be written in terms of the adjacency matrix [Thioulouse, Chessel and Champely (1995)],

$$c = \frac{(n-1)\sum_{j=1}^{n}\sum_{j=1}^{n}\mathbf{A}_{(ij)}(\mathbf{y}_{(i)} - \mathbf{y}_{(j)})^2}{2N_e\sum_i(\mathbf{y}_{(i)} - \bar{y})^2} = \frac{n-1}{N_e}\frac{\tilde{\mathbf{y}}^T(\mathbf{D}_{\mathbf{d}} - \mathbf{A})\tilde{\mathbf{y}}}{\tilde{\mathbf{y}}^T\tilde{\mathbf{y}}},$$

$$I = \frac{(n)\sum_{j=1}^{n}\sum_{j=1}^{n}\mathbf{A}_{(ij)}(\mathbf{y}_{(i)}-\bar{y})(\mathbf{y}_{(j)}-\bar{y})}{2N_e\sum_i(\mathbf{y}_{(i)}-\bar{y})^2} = \frac{n}{N_e}\frac{\tilde{\mathbf{y}}^T\mathbf{A}\tilde{\mathbf{y}}}{\tilde{\mathbf{y}}^T\tilde{\mathbf{y}}},$$

where $\tilde{\mathbf{y}} = \mathbf{y} - \bar{y}\mathbf{1}_n$ is $\mathbf{y}$ centered by the standard (unweighted) mean of the elements of $\mathbf{y}$ and $N_e = \sum_{ij}\mathbf{A}_{ij}$ is twice the number of edges in the graph. In particular, we see that Geary's c can be written in terms of an inner-product using the Laplacian.

Thioulouse, Chessel and Champely (1995) note that the variance of $\mathbf{y}$ with observations weighted by their node-degree, given by $\mathrm{var}_{\mathbf{D_d}}(\mathbf{y}) = \tilde{\mathbf{y}}^T\mathbf{D_d}\tilde{\mathbf{y}}/N_e$, can be decomposed into related components,

$$\mathrm{var}_{\mathbf{D_d}}(\mathbf{y}) = \tilde{\mathbf{y}}^T\frac{\mathbf{D_d}-\mathbf{A}}{N_e}\tilde{\mathbf{y}} + \tilde{\mathbf{y}}^T\frac{\mathbf{A}}{N_e}\tilde{\mathbf{y}}.$$

Thus, Geary's c and Moran's I are similar to F measures described above, that is, the ratio of component variability to total variability (note, however, that Moran's I can be negative). Several authors have proposed spatial multivariate analyses which rely on $\mathbf{L}$ as a metric for the rows, or a row standardized version $\mathbf{L}^* = \mathbf{D_d}^{-1}(\mathbf{D_d} - \mathbf{A})$ [Aluja-Ganet and Nonell-Torrent (1991); Thioulouse, Chessel and Champely (1995); di Bella and Jona-Lasinio (1996); Dray, Saïd and Debias (2008)]. [Note that the matrix $\mathbf{L}^*$ as well as the similar matrix $\widetilde{\mathbf{L}} = \mathbf{D_d}^{-1/2}(\mathbf{D_d} - \mathbf{A})\mathbf{D_d}^{-1/2}$ are also considered in graph theory; see Biyikoğlu, Leydold and Stadler (2007).]

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 seemingly disparate analyses as well as further exploration as to the effect of our choice of metrics.

The use of nonstandard 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.

APPENDIX A: DPCOA AND GPCA

We state here the equivalence between DPCoA and gPCA described in Section 3.3. First we describe more explicitly DPCoA, as described in Pavoine, Dufour and Chessel (2004).

DPCoA. Assume that the *squared* pairwise distances/dissimilarities between the species are given by a $S \times S$ matrix $\boldsymbol{\delta}$. We also assume that the distances are Euclidean (i.e., coordinates can be found for the points so that the standard Euclidean distance between points is given by the square-root of the entries of $\boldsymbol{\delta}$).

Following the notation provided in Section 4, let $\mathbf{P}_{\mathbf{w}_m} = (\mathbf{I}_m - \mathbf{1}_m \mathbf{w}_m^T)$ be the projection matrix that centers m observations based on a weighted mean with $\mathbf{w}_m$ as weights:

1. Find Euclidean coordinates of the species using a weighted version of Multidiminsional Scaling, with weights for the species given by $\mathbf{w}_S$, typically [and as proposed by Pavoine, Dufour and Chessel (2004)] the relative abundance of the species in all the samples. Specifically, let $\mathbf{U}$ be the eigenvectors of
$$\mathbf{D}_{\mathbf{w}_S}^{1/2} \mathbf{P}_{\mathbf{w}_S} (-\boldsymbol{\delta}/2) \mathbf{P}_{\mathbf{w}_S}^T \mathbf{D}_{\mathbf{w}_S}^{1/2}.$$
Then the new coordinates of the species are given by the rows of $\mathbf{Z} \in \mathbb{R}^{S \times s^*}$ ($s^* \leq S - 1$ is the dimension of the space required to contain the species). Then we have $\mathbf{Z} = \mathbf{D}_{\mathbf{w}_S}^{-1/2} \mathbf{U} \boldsymbol{\Lambda}^{1/2}$. Note that we could also start with a similarity matrix between species, $\mathbf{S}_\mathbf{v} = \mathbf{1}\mathbf{v}^T + \mathbf{v}\mathbf{1}^T - \frac{1}{2}\boldsymbol{\delta}$ for any $\mathbf{v}$ that implies $\mathbf{S}_\mathbf{v}$ is positive definite. Because
$$\mathbf{P}_\mathbf{w} \mathbf{S}_\mathbf{v} \mathbf{P}_\mathbf{w}^T = \mathbf{P}_\mathbf{w} (-\boldsymbol{\delta}/2) \mathbf{P}_\mathbf{w}^T$$
for any weights $\mathbf{w}$ and vector $\mathbf{v}$ the MDS will be equivalent. This is, of course, the standard equivalence between starting with a similarity matrix or dissimiliarity matrix in MDS.
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 $\mathbf{x}_\ell$). Let the rows of the $L \times s^*$ matrix $\mathbf{Y}$ contain the coordinates of the sites, so
$$\mathbf{Y} = \mathbf{X}\mathbf{Z}.$$
The squared pairwise Euclidean distance between the locations using these coordinates will be equal to their Rao Dissimilarity using the dissimilarity matrix $\boldsymbol{\delta}$.
3. Find a lower-dimensional representation of the locations using a generalized principal components analysis on the triplet $(\mathbf{Y}, \mathbf{I}_S, \mathbf{D}_{\mathbf{w}_L})$, where $\mathbf{D}_{\mathbf{w}_L}$ is a diagonal matrix consisting of weights for the locations, $\mathbf{w}_L$ (again, typically the relative abundance of the locations in all the samples). Let $r = \operatorname{rank}(\mathbf{Y})$. Then gPCA of $(\mathbf{Y}, \mathbf{I}_S, \mathbf{D}_{\mathbf{w}_L})$ gives the eigenvalue equations,
$$\mathbf{Y}^T \mathbf{D}_{\mathbf{w}_L} \mathbf{Y}\mathbf{F} = \mathbf{F}\boldsymbol{\Phi}, \qquad \mathbf{Y}\mathbf{Y}^T \mathbf{D}_{\mathbf{w}_L} \mathbf{G} = \mathbf{G}\boldsymbol{\Phi}, \tag{1}$$
$$\text{where } \mathbf{F}^T\mathbf{F} = \mathbf{I}_r, \ \mathbf{G}^T \mathbf{D}_{\mathbf{w}_L} \mathbf{G} = \mathbf{I}_r$$

and $\mathbf{Y} = \mathbf{G}\boldsymbol{\Phi}^{1/2}\mathbf{F}^T$ is the generalized SVD decomposition of $\mathbf{Y}$. The final coordinates of the locations are given by

$$\mathbf{L} = \mathbf{Y}\mathbf{F}.$$

We also transform the coordinates of the species to get species coordinates (see Section 4.2),

$$\mathbf{K} = \mathbf{Z}\mathbf{F}.$$

LEMMA. *The coordinates for the locations given by* $\mathbf{L}$ *in DPCoA using* $\boldsymbol{\delta}$ *are equivalent to the coordinates* $\widehat{\mathbf{X}} = \widetilde{\mathbf{X}}\mathbf{S}_\mathbf{v}\mathbf{A}$ *of the locations given by gPCA with the triplet* $(\widetilde{\mathbf{X}}, \mathbf{S}_\mathbf{v}, \mathbf{D}_{\mathbf{w}_L})$, *where* $\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{P}_{\mathbf{w}_S}$ *is the column centered matrix of data. 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* $\mathbf{e}_s$ *by the transformation implied from the gPCA of* $(\widetilde{\mathbf{X}}, \mathbf{S}_\mathbf{v}, \mathbf{D}_{\mathbf{w}_L})$ *so that* $\mathbf{K} = \mathbf{P}_{\mathbf{w}_S}\mathbf{S}_\mathbf{v}\mathbf{A}$.

PROOF. The fundamental eigenequations for a gPCA of the triplet $(\widetilde{\mathbf{X}}, \mathbf{S}_\mathbf{v}, \mathbf{D}_{\mathbf{w}_L})$ are

$$\begin{aligned} &\widetilde{\mathbf{X}}^T\mathbf{D}_{\mathbf{w}_L}\widetilde{\mathbf{X}}\mathbf{S}_\mathbf{v}\mathbf{A} = \mathbf{A}\boldsymbol{\Psi}, \qquad \widetilde{\mathbf{X}}\mathbf{S}_\mathbf{v}\widetilde{\mathbf{X}}^T\mathbf{D}_{\mathbf{w}_L}\mathbf{B} = \mathbf{B}\boldsymbol{\Psi}, \\ &\qquad\qquad\qquad \text{where } \mathbf{A}^T\mathbf{S}_\mathbf{v}\mathbf{A} = \mathbf{I}_r,\ \mathbf{B}^T\mathbf{D}_{\mathbf{w}_L}\mathbf{B} = \mathbf{I}_r, \end{aligned} \tag{2}$$

so that $\mathbf{X}\mathbf{P}_{\mathbf{w}_S} = \mathbf{B}\boldsymbol{\Psi}^{1/2}\mathbf{A}^T$ is the corresponding gSVD.

Since $\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{P}_{\mathbf{w}_S}$, we see that $\mathbf{B}$ and $\mathbf{G}$ from DPCoA are both eigenvectors for the same matrix, $\mathbf{X}\mathbf{P}_{\mathbf{w}_S}\mathbf{S}_\mathbf{v}\mathbf{P}_{\mathbf{w}_S}^T\mathbf{X}^T\mathbf{D}_{\mathbf{w}_L}$, implying that $\mathbf{B}$ and $\mathbf{G}$ are the $\mathbf{D}_{\mathbf{w}_L}$-orthonormal eigenvectors of the same matrix. This implies that the eigenvalues are the same ($\boldsymbol{\Phi} = \boldsymbol{\Psi}$) 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 $\mathbf{L} = \mathbf{Y}\mathbf{F} = \mathbf{G}\boldsymbol{\Phi}^{1/2}$. With gPCA of $(\widetilde{\mathbf{X}}, \mathbf{S}_\mathbf{v}, \mathbf{D}_{\mathbf{w}_L})$, the location coordinates are $\widehat{\mathbf{X}} = \mathbf{X}\mathbf{P}_{\mathbf{w}_S}\mathbf{S}_\mathbf{v}\mathbf{A} = \mathbf{B}\boldsymbol{\Psi}^{1/2}$ and, therefore, we have that $\mathbf{L} = \widehat{\mathbf{X}}$—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 $\mathbf{F}^T = \boldsymbol{\Phi}^{-1/2}\mathbf{G}^T\mathbf{D}_{\mathbf{w}_L}\mathbf{Y}$ and, similarly, $\mathbf{B}\boldsymbol{\Psi}^{-1/2} = \mathbf{X}\mathbf{P}_{\mathbf{w}_S}\mathbf{S}_\mathbf{v}\mathbf{A}\boldsymbol{\Psi}^{-1}$. Remembering that $\mathbf{Z}\mathbf{Z}^T = \mathbf{P}_{\mathbf{w}_S}\boldsymbol{\delta}\mathbf{P}_{\mathbf{w}_S}^T$, the final coordinates of the species from DPCoA are given by

$$\begin{aligned} \mathbf{K} &= \mathbf{Z}\mathbf{Y}^T\mathbf{D}_{\mathbf{w}_L}\mathbf{G}\boldsymbol{\Phi}^{-1/2} = \mathbf{Z}\mathbf{Z}^T\mathbf{X}^T\mathbf{D}_{\mathbf{w}_L}\mathbf{G}\boldsymbol{\Phi}^{-1/2} \\ &= \mathbf{P}_{\mathbf{w}_S}\boldsymbol{\delta}\mathbf{P}_{\mathbf{w}_S}^T\mathbf{X}^T\mathbf{D}_{\mathbf{w}_L}\mathbf{G}\boldsymbol{\Phi}^{-1/2} \\ &= \mathbf{P}_{\mathbf{w}_S}\mathbf{S}_\mathbf{v}\underbrace{\mathbf{P}_{\mathbf{w}_S}^T\mathbf{X}^T\mathbf{D}_{\mathbf{w}_L}\mathbf{X}\mathbf{P}_{\mathbf{w}_S}\mathbf{S}_\mathbf{v}\mathbf{A}}_{=\mathbf{A}\boldsymbol{\Psi}\text{ from (2)}}\boldsymbol{\Psi}^{-1} = \mathbf{P}_{\mathbf{w}_S}\mathbf{S}_\mathbf{v}\mathbf{A} \end{aligned}$$

up to the sign change difference between $\mathbf{G}$ and $\mathbf{B}$. □

APPENDIX B: KERNEL ANALYSIS AND GPCA

Multivariate kernel methods seek a set of functions $f_1, \dots, f_k$ from our general data space $\mathcal{X}$ into $\mathbb{R}$, such that the possible set of functions f form a Reproducing Kernel Hilbert Space with respect to a kernel function $\mathcal{K}$ on $\mathcal{X}$ [see Schölkopf and Smola (2002) for details]. The solutions for a multivariate Kernel CCA (or extensions) can be recovered from the eigenequations, assuming $\mathbf{K}_i$ are invertible

$$\mathbf{K}_{\xi_2}^{-1}\mathbf{K}_2\mathbf{K}_1\mathbf{U}_1 = \mathbf{K}_{\xi_1}\mathbf{U}_1\mathbf{\Lambda}$$

with the constraint that

$$\mathbf{U}_i^T\mathbf{K}_{\xi_i}\mathbf{U}_i = \mathbf{\Gamma}_i$$

and $\mathbf{U}_2$ is given by

$$\mathbf{U}_2 = \mathbf{K}_{\xi_2}^{-1}\mathbf{K}_1\mathbf{U}_1(\mathbf{\Lambda}\mathbf{\Gamma}_1\mathbf{\Gamma}_2^{-1})^{-1/2},$$

where $\mathbf{K}_{\xi_i} = (1-\xi_i)\mathbf{K}_i/n + \xi_i\mathbf{I}$, and $\mathbf{\Gamma}_i$ are diagonals of normalization constants chosen by the user. Then the new coordinates of an object x from data set i are given by $(f_1(x), \dots, f_k(x))^T = \mathbf{k}^T\mathbf{U}_i$, where $\mathbf{k}_{(j)} = \mathcal{K}_i(x, x_j)$.

Let $\mathbf{K}_1 = \mathbf{Q}_n$, $\mathbf{K}_2 = \mathbf{X}\mathbf{Q}_p\mathbf{X}^T$, and $\xi_1 = \xi_2 = 1$, then we have the eigenequation

$$\mathbf{X}\mathbf{Q}_p\mathbf{X}^T\mathbf{Q}_n\mathbf{U}_1 = \mathbf{U}_1\mathbf{\Lambda}. \tag{3}$$

We see that these are equivalent to the gPCA equations, with $\mathbf{U_1} = \mathbf{B}$. Then the coordinates associated with the data matrix $\mathbf{X}$ from the kernel method are

$$\mathbf{K}_2\mathbf{U}_2 = \mathbf{K}_2\mathbf{K}_1\mathbf{U}_1(\mathbf{\Lambda}\mathbf{\Gamma}_1\mathbf{\Gamma}_2^{-1})^{-1/2} = \mathbf{X}\mathbf{Q}_p\mathbf{X}^T\mathbf{Q}_n\mathbf{U}_1\mathbf{\Lambda}^{-1/2}(\mathbf{\Gamma}_1\mathbf{\Gamma}_2^{-1})^{-1/2},$$

while those from the gPCA are

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

Choosing the scaling of the eigenvectors so that $(\mathbf{\Gamma}_1\mathbf{\Gamma}_2^{-1}) = \mathbf{I}$ makes the solutions equivalent.

APPENDIX C: INERTIA AND DISSIMILARIES

We generalize the results of Pélissier et al. (2003) to show the derivation of the dissimilarity and diversity results above.

In gPCA, the term *inertia* is used for the inter-point similarities, and the total inertia between points is defined as $I(\mathbf{X}, \mathbf{Q}_p, \mathbf{Q}_n) = \operatorname{tr}(\mathbf{Q}_n\mathbf{X}\mathbf{Q}_p\mathbf{X}^T) = \sum\lambda_i$. Then if $\widehat{\mathbf{X}}_{(r)}$ are the new coordinates of $\mathbf{X}$ restricted to the first r dimensions and $\widehat{\mathbf{X}}_{(-r)}$ the remaining $p-r$ dimensions, we can decompose the total inertia into the inertia of the first r dimensions and that of the remaining $p-r$,

$$\begin{aligned} I(\mathbf{X}, \mathbf{Q}_p, \mathbf{Q}_n) &= I(\widehat{\mathbf{X}}_{(r)}, \mathbf{I}_p, \mathbf{Q}_n) + I(\widehat{\mathbf{X}}_{(-r)}, \mathbf{I}_p, \mathbf{Q}_n) \\ &= \sum_{i=1}^{r}\lambda_i + \sum_{i=r+1}^{p}\lambda_i, \end{aligned}$$

and the first r dimensions give maximal possible inertia for r dimensions.

Let $\mathbf{Y} \in \mathbb{R}^{N\times S}$ be the incidence matrix for the species variable, where $\mathbf{Y}_{(is)}$ is an indicator of the ith observation being species s. Let $\mathbf{Z} \in \mathbb{R}^{N\times L}$ be a similar such incidence matrix for the location variable. Then the inertia of the eigenanalysis of the triplet $(\widetilde{\mathbf{Y}}, \mathbf{Q}, \mathbf{D}_N)$, where $\widetilde{\mathbf{Y}}$ is the (nonweighted) centered $\mathbf{Y}$ and $\mathbf{D}_N$ is a diagonal matrix of N elements, will be equal to $H_{\mathbf{Q}}(\bar{\mathbf{x}})$. Regressing $\mathbf{Y}$ onto $\mathbf{Z}$ gives predictions $\widetilde{\mathbf{Y}}_Z$ and residuals $\widetilde{\mathbf{Y}}_{|Z} = \widetilde{\mathbf{Y}} - \widetilde{\mathbf{Y}}_Z$. Then the total inertia (I_{T}) can be broken into the inertia due to differences between locations (I_{B}) plus the remaining inertia within locations (I_{W}),

$$\mathrm{Inertia}(\widetilde{\mathbf{Y}}, \mathbf{Q}, \mathbf{D}_N) = \mathrm{Inertia}(\widetilde{\mathbf{Y}}_Z, \mathbf{Q}, \mathbf{D}_N) + \mathrm{Inertia}(\widetilde{\mathbf{Y}}_{|Z}, \mathbf{Q}, \mathbf{D}_N).$$

Note that $\widetilde{\mathbf{Y}}_Z = \mathbf{Z}\widetilde{\mathbf{X}}$, so that the inertia explained by $\mathbf{Z}$ is equal to the inertia of the eigenanalysis of $\widetilde{\mathbf{X}}$,

$$I_{\mathrm{B}} = \mathrm{Inertia}(\widetilde{\mathbf{Y}}_Z, \mathbf{Q}, \mathbf{D}_N) = \mathrm{Inertia}(\widetilde{\mathbf{X}}, \mathbf{Q}, \mathbf{D}_{\mathbf{w}_L}).$$

This implies that the ordination procedures described above best preserve the between location dissimilarities defined by the metric $\mathbf{Q}$.

APPENDIX D: 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 nonzero components are the same sign distinct regions of the graph. Those eigenvectors corresponding to large eigenvalues tend 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 $\mathbf{x}$ associated with a graph, with each element of $\mathbf{x}$ 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, Kirkland and Neumann (2005)]. Let $\boldsymbol{\delta}_T$ 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

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

where for a phylogenetic tree $\mathbf{v}$ 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

$$\mathbf{L}_S^{-1} = c\boldsymbol{\gamma}\boldsymbol{\gamma}^T - 1/2\boldsymbol{\delta}_S, \qquad \text{where } c = (8\mathbf{1}^T\boldsymbol{\delta}_{S\times I}\mathbf{1})^{-1},\ \boldsymbol{\gamma} = \boldsymbol{\delta}_T\mathbf{v},$$

and $\boldsymbol{\delta}_S \in \mathbb{R}^{S\times S}$ is the distance matrix restricted to the distances between leaves of the tree and $\boldsymbol{\delta}_{SI} \in \mathbb{R}^{S\times S-1}$ 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 $\mathbf{P}\boldsymbol{\gamma}\boldsymbol{\gamma}\mathbf{P}^T$ 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 $\boldsymbol{\Sigma}$.

APPENDIX E: EIGENVECTORS OF $\boldsymbol{\Sigma}$ FOR A PHYLOGENETIC TREE

Note the block structure in $\boldsymbol{\Sigma}$: 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 $\boldsymbol{\Sigma}$ so that

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \varnothing \\ \varnothing & \boldsymbol{\Sigma}_2 \end{pmatrix}, \tag{4}$$

where $\boldsymbol{\Sigma}_1$ is a $S_1 \times S_1$ matrix, S_1 is the number of extant species descended from $\mathcal{P}_1$, and similarly with $\boldsymbol{\Sigma}_2$. This means that the eigenvectors of $\boldsymbol{\Sigma}$ must be of the form

$$\begin{pmatrix} \mathbf{v}_{1i} \\ \varnothing \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} \varnothing \\ \mathbf{v}_{2j} \end{pmatrix}, \tag{5}$$

where $\{\mathbf{v}_{1i}\}_{i=1}^{S_1}$ are the eigenvectors of $\mathbf{\Sigma}_1$ and $\{\mathbf{v}_{2j}\}_{j=1}^{S_2}$ are the eigenvectors of $\mathbf{\Sigma}_2$. Therefore, every eigenvector of $\mathbf{\Sigma}$, at a minimum, must be only nonzero for one of the lineages.

Indeed, if we think back to the definition of $\mathbf{\Sigma}$, the elements of the blocks $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_2$ are themselves rank-1 perturbations of block diagonal matrices:

$$\mathbf{\Sigma}_1 = \begin{pmatrix} \mathbf{\Sigma}_{11} & \varnothing \\ \varnothing & \mathbf{\Sigma}_{12} \end{pmatrix} + c_1 \mathbf{1}\mathbf{1}^T, \qquad \mathbf{\Sigma}_2 = \begin{pmatrix} \mathbf{\Sigma}_{21} & \varnothing \\ \varnothing & \mathbf{\Sigma}_{22} \end{pmatrix} + c_2 \mathbf{1}\mathbf{1}^T, \tag{6}$$

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

The claims in the literature for a relationship of the eigenvectors of $\mathbf{\Sigma}$ 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 ($\mathbf{t} \propto \mathbf{1}$). 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 $\mathbf{\Sigma}$ 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 $\mathbf{\Sigma}$, at a minimum, must consist of zero elements for one branch of the tree.

Beyond this, Cavalli-Sforza and Piazza (1975) describe "usual" behavior of the eigenvectors, but their ideas do not scale as the size of the tree increases. The nested block structure of $\mathbf{\Sigma}$ 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.

APPENDIX F: ELLIPSES IN DPCOA PLOTS

The ellipse plots given in Figure 2(b) are provided by the ade4 package and represent a location vector, $\mathbf{x}_\ell$, as an ellipse. For completeness, we explain here what ade4 is plotting.

Let the *species* coordinates as transformed into two dimensions by the ordination of the *locations* be given in the columns of $\mathbf{K}_2 \in \mathbb{R}^{S\times 2}$. Then the ellipsoid for $\mathbf{x}_\ell \in \mathbb{R}^S$ is defined by

$$\mathbf{v}^T(\mathbf{K}_2^T \mathbf{D}_{\mathbf{x}_l} \mathbf{K}_2)^{-1}\mathbf{v} = 1,$$

where the ellipse is centered at $\mathbf{x}_l$.

This curve consists of the points with norm 1 in the Mahalanobis metric, only the estimate of the variance in Mahalanobis distance is calculated with weights on the points (species) given by $\mathbf{x}_l$. Equivalently, the ellipses in Figure 2(b) will have major and minor axes in the direction of the weighted principal components of the coordinates of the species in $\mathbf{K}_2$, with the lengths of the axes given by the weighted standard deviation of the species coordinates in those directions (an ellipse defined by the equation $\mathbf{x}^T\mathbf{Q}\mathbf{x} = 1$ will have major and minor axes in the directions of the eigenvectors of $\mathbf{Q}$ with lengths given by $1/\sqrt{\lambda_i}$, where λ_i is an eigenvalue of $\mathbf{Q}$).

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

REFERENCES

Aluja-Ganet, T. and Nonell-Torrent, R. (1991). Local principal components analysis. *Questiio* **15** 267–278.

Bach, F. R. and Jordan, M. I. (2002). Kernel independent component analysis. *J. Mach. Learn. Res.* **3** 1–48. MR1966051

Bapat, R., Kirkland, S. J. and Neumann, M. (2005). On distance matrices and Laplacians. *Linear Algebra Appl.* **401** 193–209. MR2133282

Biyikoğlu, T., Leydold, J. and Stadler, P. F. (2007). *Laplacian Eigenvectors of Graphs. Lecture Notes in Mathematics* **1915**. Springer, Berlin. MR2340484

Cavalli-Sforza, L. L. and Piazza, A. (1975). Analysis of evolution: Evolutionary rates, independence and treeness. *Theoretical Population Biology* **8** 127–165. MR0526635

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.

D'Ambra, L. and Lauro, N. C. (1992). Non-symmetrical exploratory data analysis. *Statist. Appl.* **4** 511–529.

di Bella, G. and Jona-Lasinio, G. (1996). Including spatial contiguity information in the analysis of multispecific patterns. *Environmental and Ecological Statistics* **3** 260–280.

Diestel, R. (2005). *Graph Theory*, 3rd ed. *Graduate Texts in Mathematics* **173**. Springer, New York. MR2159259

Dray, S. and Dufour, A.-B. (2007). The ade4 package: Implementing the duality diagram for ecologists. *J. Statist. Softw.* **22**.

Dray, S., Saïd, S. and Debias, F. (2008). Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. *Journal of Vegetation Science* **19** 45–56.

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. Legendre, eds.). *NATO ASI Series* **G14** 139–156. Springer, Berlin. MR0913539

EXCOFFIER, L., SMOUSE, P. and QUATTRO, J. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics* **131** 479–491.

FELSENSTEIN, J. (1981). Evolutionary trees from gene frequencies and quantitative characters: Finding maximum likelihood estimates. *Evolution* **35** 1229–1242.

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. Johns Hopkins Univ. Press, Baltimore. MR1417720

GREENACRE, M. J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, London. MR0767260

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. (2008). Multivariate analysis: The French way. In *Probability and Statistics*: *Essays in Honor of David A. Freedman* (D. Nolan and T. Speed, eds.). *IMS Lecture Notes* **2** 219–233. IMS, Beachwood, OH. MR2459953

JOLLIFFE, I. T. (2002). *Principal Components Analysis*, 2nd ed. Springer, New York. MR2036084

KONDOR, R. I. and LAFFERTY, J. (2002). Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of ICML* 315–322.

LEGENDRE, P. and LEGENDRE, L. (1998). *Numerical Ecology*, 2nd English ed. *Developments in Environmental Modeling* **20**. Elsevier, New York. MR1675780

MAESSCHALCK, R. D., JOUAN-RIMBAUD, D. and MASSART, D. (2000). The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems* **50** 1–18.

MARTIN, A. (2002). Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Applied and Environmental Microbiology* **68** 3673–3682.

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. *J. Theoret. Biol.* **228** 523–537. MR2080909

PAVOINE, S., OLLIER, S., PONTIER, D. and CHESSEL, D. (2008). Testing for phylogenetic signal in phenotypic traits: New matrices of phylogenetic proximities. *Theoretical Population Biology* **73** 79–91.

PÉLISSIER, R., COUTERON, P., DRAY, S. and SABATIER, D. (2003). Consistency between ordination techniques and diversity measurements: Two strategies for species occurrence data. *Ecology* **84** 242–251.

PURDOM, E. (2006). Multivariate kernel methods in the analysis of graphical structures. Ph.D. thesis, Stanford Univ. MR2709407

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.

RAO, C. R. (1982). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology* **21** 24–43. MR0662520

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.

SCHÖLKOPF, B. and SMOLA, A. J. (2002). *Learning with Kernels*: *Support Vector Machines*, *Regularization*, *Optimization*, *and Beyond*. MIT Press, Cambridge, MA.

THIOULOUSE, J., CHESSEL, D. and CHAMPELY, S. (1995). Multivariate analysis of spatial patterns: A unified approach to local and global structures. *Environmental and Ecological Statistics* **2** 1–14.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA AT BERKELEY
367 EVANS HALL #3860
BERKELEY, CALIFORNIA 94720-3860
USA
E-MAIL: epurdom@stat.berkeley.edu