Determining Significance in Metagenomic Samples

Gregory Ditzler, Robi Polikar and Gail Rosen

***Abstract*—Many ecology studies employ ordination methods to visually inspect metagenomic data sets, which initially may contain thousands of dimensions that represent operational taxonomic units (OTUs) of a sample. Many times, MANOVA (applied to a pairwise distance set) is applied to determine how different the groups in the study are from one another. It is convenient to have a p-value that allows us to interpret if two or more groups are different than one another with statistical confidence, where the null hypothesis is that the two populations are not different than the other. With MANOVA all groups are tested under the hypothesis that they are equal. In this work, we present a statistical framework for obtaining a p-value to compare multiple groups that is derived from a non-parametric statistical test, which uses data derived from the OTU features. The result is a matrix of p-values for the comparison on multiple groups in a metagenomic data set. We test our approach on a real-world database using several variations of ordination techniques.**

I. Introduction

Metagenomics is the study uncultured microorganisms obtained directly from an environmental sample [1]. In ecology, scientists are not only concerned about what species are in an environmental sample, but also how different samples compare [2]. Through next generation sequencing [3], it is now possible to collect, process and annotate sequences obtained from a microbe that contains thousands of microbial species, which may provide a plethora of information about the site from where the sample was obtained. Related efforts in the study of the human microbiome have traditionally used standard coordinate analysis schemes with little justification of the methods used in the analysis. In this work we use several coordinate analysis schemes along with a new implementation that takes advantage of kernels for measuring distance in a feature space. We also present a quick method that tests for significance between groups in a metagenomic sample that uses data collected from PCoA. The test for significance is computed using pairwise comparisons of groups in a data set, which returns multiple p-values for a group-by-group comparison.

II. Methods

A. Measuring Significance

In this work we present a method to compute a p-value that measures the whether probability of two groups of samples are indeed different from each other using a non-parametric approach. The pseudo code is highlighted in Fig. 1. To begin, we separate the dataset, $\mathcal{D}$, into C different groups where C is the number of categories in a metagenomic set of samples (e.g., gut, oral, or skin samples). Each sample is represented by a d-dimensional vector, which is most likely the product of a pre-processing step such as principal coordinate analysis (PCoA) [4] applied to the OTU abundance table. Then, for each feature, we apply the Kolmogorov-Smirnov (KS) test to each pairwise combination of categories in the data set. Thus, the result is a $d \times 1$ vector containing the p-values for each feature comparison of class i and j computed via the KS-test, where i and j are two arbitrary categories (i.e., classes) in the data set. Fisher's method is applied to combine the d p-values into a single quantity, X^2, which is given by (1). Then, X^2 is compared to a χ^2-distribution with $2d$ degrees of freedom to obtain a p-value. This p-value may then be used for hypothesis testing. Using this technique, we obtain a $C \times C$ matrix, $\mathbf{P}$, containing the pairwise comparison between all categories.

```
Input: Set $\mathcal{D} = \{(\vec{x}_n, y_n)\}_{n=1}^{m}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = [C]$
for $i = 1 \ldots, C$ do
    for $j = 1 \ldots, C$ do
        Initialize: $\vec{\rho}$ to a $d \times 1$ vector
        for $k = 1, \ldots, d$ do
            Apply KS-test to $\mathcal{D}_i^k$ and $\mathcal{D}_j^k$ where $\mathcal{D}_i^k$ and $\mathcal{D}_j^k$ are
            the $k$-th feature of data from categories $i$ and $j$ in $\mathcal{D}$,
            respectively. Store the $p$-value from the KS-test in $\vec{\rho}_k$
        end for
        $X^2 = -2 \sum_{k=1}^{d} \log \vec{\rho}_k$    (1)
        $X^2$ is distributed as a $\chi^2$ distribution, then set the $p$-value
        for $\{\mathbf{P}\}_{ij}$ by comparing $X^2$ to a $\chi^2$ distribution with $2d$
        degrees of freedom
    end for
end for
```

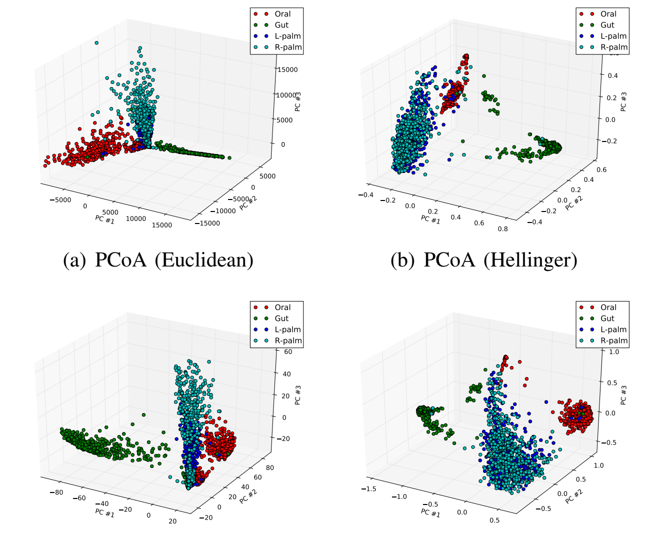
Fig. 1. Pseudo code for computing differences between categories in a metagenomic sample

B. New Methods for Ordination

Kernel based methods have shown great success in many areas of machine learning including classification, regression and component analysis [5]. It only seems natural to apply kernel methods to PCoA techniques, as PCoA is used by many biologists and ecologists. We use kernels as distance measures, which are computed as norms in feature space. Traditional positive definite (pd) kernels provide us with a

G. Ditzler and G. Rosen are with the EESI Laboratory in the ECE Department at Drexel University. They are supported by the NSF CAREER award #0845827, NSF Award #1120622, and the Dept. of Energy Award #SC004335. Author email: gregory.ditzler@gmail.com and gailr@ece.drexel.edu

R. Polikar is with the SPPRL in the ECE Department at Rowan University. He is supported by the NSF under Grant No: ECCS-0926159. Author email: polikar@rowan.edu



(a) PCoA (Euclidean)

(b) PCoA (Hellinger)

(c) KPCoA (MQ)

(d) NMDS

Fig. 2. PCoA applied to the 15-month study of the human microbiome data set with a variation in the selection of the distance method.

measure of similarity by way of the canonical dot product; however, a large class of kernels, known as conditionally positive definite (cpd), also exist for measuring dissimilarity rather than similarity in feature space. For example, a norm in feature space is calculated as:

$$\|\Phi(\vec{x}) - \Phi(\vec{x}')\|^2 = k(\vec{x}, \vec{x}) + k(\vec{x}', \vec{x}') - 2k(\vec{x}, \vec{x}') \qquad (2)$$

where $k(\vec{x}, \vec{x}') = \Phi(\vec{x})^\mathsf{T}\Phi(\vec{x}')$ is the kernel and $\Phi(\vec{x})$ is a non-linear function applied to $\vec{x}$. While there is a large body of work that describes the properties of cpd kernels, we simply note that the pairwise distance matrix used in PCoA can be computed using kernels that measure a norm distance in feature space. For this work we have selected the multi-quadratic kernel for the experiments as described in [5] for PCoA (we refer to this method as KPCoA).

C. Implementation Overview

We use the 15-month study of the human microbiome as the basis for testing our method [6]. The data set is collected from two subjects over a 15 month period from four different body sites (oral cavity, gut, left palm and right palm). The PyCogent[1] toolbox was used for the implementation of the PCoA methods. The significance procedure described above is applied to the data collected after PCoA.

III. Results

TABLE I
THE MATRIX $\mathbf{P}$ FOR PCoA WITH THE HELLINGER DISTANCE ON THE HUMAN MICROBIOME DATA.

	Gut	Oral	L-Palm	R-Palm
Gut	1.0	0.0	0.0	0.0
Oral	0.0	1.0	0.0	0.0
L-Palm	0.0	0.0	1.0	0.875
R-Palm	0.0	0.0	0.875	1.0

The PCoA plots for the human microbiome study are shown in Fig. 2. Results from PCoA computed using the Euclidean/Hellinger distance, KPCoA with the multi-quadratic kernel, and non-metric multidimensional scaling are presented here. Clearly, the selection of the distance measure used during the analysis is capable of proving or disproving a hypothesis given the same set of data. Table I shows the p-values computed using our approach (recall that a low p-value, say less then 0.05 rejects the null hypothesis that the two sets of samples come from the same distribution). The table shows that there is not enough evidence to reject the null hypothesis for the samples collected from the skin at a 95% confidence level; however, there is sufficient evidence to reject the hypothesis for gut-oral, gut-skin and oral-skin comparisons. Unfortunately, selecting a different distance measure to compare samples can potentially provide different significance results as is the case with Euclidean PCoA and KPCoA. Hence, using PCoA in an attempt to inspect whether multiple groups are different may be affected by selected distance measure used in the analysis. Furthermore, we performed permutation based MANOVA to the distance matrix for PCoA. From this analysis, we find significant differences between the samples at 99% confidence; however, using MANOVA in this manner does not allow us to observe which groups are significantly different from others.

IV. Conclusion

In this work we have presented a straightforward approach to measure statistical significance between populations in metagenomic data after PCoA was applied. Furthermore, we have demonstrated that care needs to be taken when multiple populations are compared using data derived from PCoA, because the selection of the distance measure is capable of proving or disproving a hypothesis. Finally, we note that since our approach uses the KS-test, we do not need to make any assumption about the distribution of data (e.g., many commonly used methods like ANOVA assume a normal model).

References

[1] J. C. Wooley, A. Godzik, and I. Friedberg, "A primer on metagenomics," *PLoS Computational Biology*, vol. 6, no. 2, pp. 1–13, 2010.

[2] G. L. Rosen, B. A. Sokhansanj, R. Polikar, M. A. Bruns, J. Russell, E. Garbarine, S. Essinger, and N. Yok, "Signal processing for metagenomics: Extracting information from the soup," *Current Genomics*, vol. 10, no. 7, pp. 493–510, 2009.

[3] M. L. Metzker, "Sequencing technologies – the next generation," *Nature Methods*, vol. 11, pp. 31–46, 2010.

[4] J. C. Gower, "Multivariate analysis and multidimensional geometry," *Journal of the Royal Statistical Society*, vol. 17, no. 1, pp. 13–28, 1967.

[5] B. Schlköpf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 1st ed., 2001.

[6] J. G. Caporaso, C. L. Lauber, E. K. Costello, D. Berg-Lyons, A. Gonzalez, J. Stombaugh, D. Knights, P. Gajer, J. Ravel, N. Fierer, J. I. Gordon, and R. Knight, "Moving pictures of the human microbiome," *Genome Biology*, vol. 12, no. 5, 2011.

[1] http://pycogent.wordpress.com/