Bioinformatics, 36(13), 2020, 4099–4101
doi: 10.1093/bioinformatics/btaa276
Advance Access Publication Date: 27 April 2020
Applications Note

Data and text mining

aPCoA: covariate adjusted principal coordinates analysis

Yushu Shi [1], Liangliang Zhang[1], Kim-Anh Do[1], Christine B. Peterson[1,*] and Robert R. Jenq[2,*]

[1]Department of Biostatistics and [2]Department of Genomic Medicine, University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on January 16, 2020; revised on March 27, 2020; editorial decision on April 19, 2020; accepted on April 21, 2020

Abstract

Summary: In fields, such as ecology, microbiology and genomics, non-Euclidean distances are widely applied to describe pairwise dissimilarity between samples. Given these pairwise distances, principal coordinates analysis is commonly used to construct a visualization of the data. However, confounding covariates can make patterns related to the scientific question of interest difficult to observe. We provide adjusted principal coordinates analysis as an easy-to-use tool, available as both an R package and a Shiny app, to improve data visualization in this context, enabling enhanced presentation of the effects of interest.

Availability and implementation: The R package 'aPCoA' and Shiny app can be accessed at https://cran.r-project.org/web/packages/aPCoA/index.html and https://biostatistics.mdanderson.org/shinyapps/aPCoA/.

Contact: cbpeterson@mdanderson.org or rrjenq@mdanderson.org

1 Introduction

Non-Euclidean distances, such as Bray–Curtis dissimilarity (Bray and Curtis, 1957), unweighted UniFrac distance (Lozupone and Knight, 2005) and weighted UniFrac distance (Lozupone *et al.*, 2007) are widely used in fields such as ecology and microbiology to describe pairwise dissimilarity between samples. In these applications, non-Euclidean distances have critical advantages over Euclidean distances, such as handling extreme values and incorporating phylogenetic information. Given a non-Euclidean pairwise distance matrix, principal coordinates analysis (PCoA), also known as classic or metric multidimensional scaling, can allow researchers to visualize variation across samples and potentially identify clusters by projecting the observations into a lower dimension.

A long-standing challenge in PCoA visualization is that confounding covariates can mask the effect of the primary covariate. For instance, in a study on the impact of diet on the microbiome, clustering due to site may be more visually prominent than diet if patients are recruited from two different locations. Though there have been several methods proposed to adjust for covariates in principal component analysis (Chang and Du, 1999; Lin *et al.*, 2016), there are no existing methods to adjust for covariates in PCoA. In this work, we develop a novel visualization approach, adjusted principal coordinates analysis (aPCoA), which allows adjustment for covariates in creating the PCoA projection, and provide easy-to-use R tools implementing this method.

2 Materials and methods

In this section, we first review the standard steps in creating a PCoA projection from an $N \times N$ distance matrix $\mathbf{D}$ summarizing the pairwise dissimilarity among the N samples in the dataset. We then describe how we modify this approach to incorporate covariate adjustment. The standard steps for PCoA are:

1. Transform $\mathbf{D}$ to a new matrix $\mathbf{A} = [A_{bi}]$, where $a_{bi} = -1/2D_{bi}^2$.
2. Center $\mathbf{A}$ to get Gower's centered matrix $\mathbf{G} = (\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{N})\mathbf{A}(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{N})$.
3. Calculate the eigendecomposition of $\mathbf{G}$.
4. Project the N samples into two dimensions determined by the two leading eigenvectors.

If the distances are Euclidean embeddable, there exists an $N \times P$ data matrix $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_N]'$, such that Gower's centered matrix can be equivalently calculated from $\mathbf{G} = \mathbf{Y}_C\mathbf{Y}_C'$, where $\mathbf{Y}_C$ is $\mathbf{Y}$ centered by the sample mean (Gower, 1966; Legendre and Legendre, 2012). To construct the aPCoA projection, we adjust for the effect of covariates on Gower's matrix, in a manner similar to MANOVA. The S covariates we want to adjust for can be represented in a $N \times S$ matrix, $\mathbf{X} = [X_1, X_2, \ldots, X_N]'$, where each X_k' is an $1 \times S$ vector. We use a matrix $\mathbf{E}$ to denote the error term which cannot be explained after doing a linear regression on $\mathbf{X}$:

$$\mathbf{E} = (\mathbf{I} - \mathbf{H})\mathbf{Y}_C, \quad (1)$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the hat matrix used in linear regression. The error covariance matrix, which is also used in pseudo F statistics calculation (Pan, 2011; Zapala and Schork, 2006) can be calculated by:

$$\Delta = \mathbf{E}\mathbf{E}' = (\mathbf{I} - \mathbf{H})\mathbf{Y}_C\mathbf{Y}_C'(\mathbf{I} - \mathbf{H}). \quad (2)$$

For any non-Euclidean distance, if we substitute $\mathbf{Y}_C'\mathbf{Y}_C$ in (2) with the corresponding Gower's centered matrix **G**, we can get the generalized error matrix, which is also the covariate adjusted Gower's centered matrix.

$$\Delta^* = (\mathbf{I} - \mathbf{H})\mathbf{G}(\mathbf{I} - \mathbf{H}). \quad (3)$$

After calculating the eigenvectors and eigenvalues of Δ^*, we can visualize this covariate adjusted Gower's matrix as in a normal PCoA plot.

We provide aPCoA as both an R package and Shiny app. The Shiny app allows for the adjustment of one covariate, which can be either continuous or categorical, and provides options for visualization including the plotting of 95% confidence ellipses and lines linking cluster members to the cluster center. Our R package additionally enables adjustment for multiple covariates.

3 Illustrating example

The first illustrating dataset is from a study on the effects of disturbance from a soldier crab on 56 species of meiobenthos, which are small invertebrates (Wang *et al.*, 2019). Eight of the sixteen observations in the dataset correspond to crab disturbances. Besides the crab disturbance, there are also four different locations in the study design, where observations from each location comprised two disturbed and two undisturbed ones. Here, we use the Bray–Curtis dissimilarity, which is commonly used in the ecology field to visualize observations.

The second illustrating dataset is from a two-center pancreatic cancer study (Riquelme *et al.*, 2019), which includes 25 patients from one hospital and 43 patients from another hospital. The investigator compared the tumor microbiota between the 36 long-time survivors (LTS) and 32 short-time survivors (STS) across study centers. The metric used for visualization is the weighted UniFrac distance, which incorporates both the taxa abundance and phylogenetic relatedness of the bacterial taxa.

As shown in the uppermost panels of Figure 1A, the original PCoA plot of the meiobenthos dataset is affected by the location, and all locations are separated from each other. After removing the effect of location using aPCoA, the separation between the disturbed and undisturbed groups becomes more prominent, whereas the separation due to location is less apparent, as shown in the bottom panels of Figure 1A.

In the pancreatic cancer example, the original PCoA plot with weighted UniFrac distance does not clearly separate the LTS and STS patients due to the confounding effect of hospital site, as shown in the upper part of Figure 1B. After adjusting for the site effect, the two clusters become more visually separable, as shown in the aPCoA plots provided in the bottom two panels.

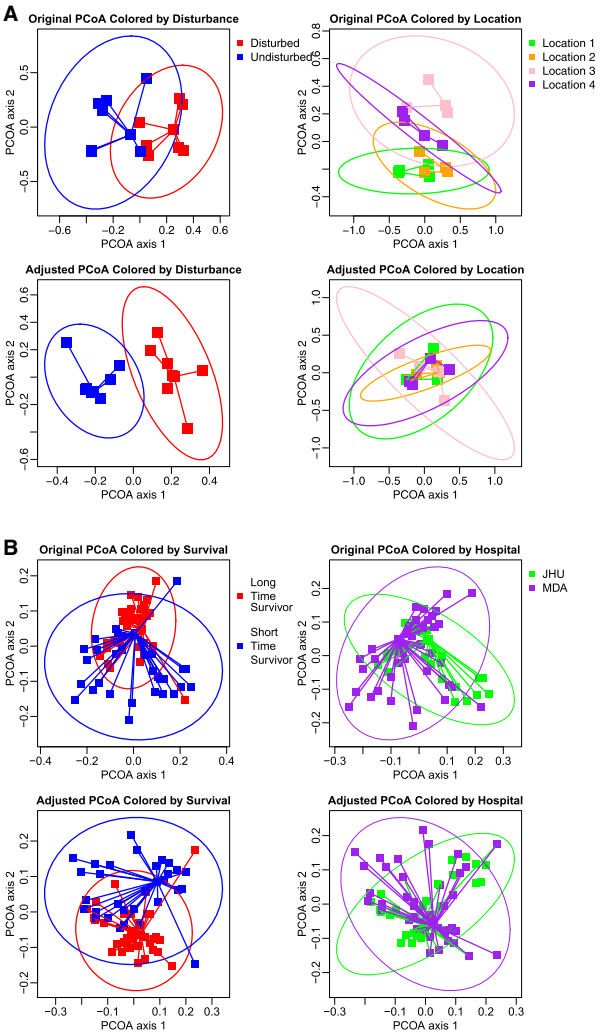

Fig. 1. Comparison between original PCoA plot and aPCoA plot for two illustrative examples. The position of the points is identical between the left and right columns, with coloring and lines used to illustrate grouping variables. (**A**) Meiobenthos dataset using the Bray-Curtis dissimilarity. (**B**) Pancreatic cancer dataset using the weighted UniFrac distance

4 Conclusion

We introduce covariate adjusted PCoA visualization along with an R implementation, which can help researchers visualize main effects in datasets with strong confounders. We expect our method to be a useful tool for microbiome and ecology researches in the future.

Funding

K.-A.D. was partially supported by the MD Anderson Moon Shot Programs, Prostate Cancer SPORE P50CA140388, NIH/NCI CCSG grant P30CA016672, CCTS 5UL1TR000371 and CPRIT RP160693 grants. C.B.P. was partially supported by the NIH/NCI CCSG grant P30CA016672 and MD Anderson Moon Shot Programs. R.R.J. was partially supported by NIH R01 HL124112 and CPRIT RR160089 grants.

Conflict of Interest: R.R.J. has consulted for Karius, Merck, Microbiome DX, and Prolacta, and is on the scientific advisory boards of Kaleido, Maat Pharma, and Seres, and has received patent royalties licensed to Seres.

References

Bray,J.R. and Curtis,J.T. (1957) An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.*, **27**, 325–349.

Chang,C.-I. and Du,Q. (1999) Interference and noise-adjusted principal components analysis. *IEEE Geosci. Remote Sens. Lett.*, **37**, 2387–2396.

Gower,J.C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325–338.

Legendre,P. and Legendre,L. (2012) Chapter 9 – ordination in reduced space. In: Pierre,L. and Louis,L. (eds) *Numerical Ecology*, *volume 24 of Developments in Environmental Modelling*. Elsevier, Oxford, UK, pp. 425–520.

Lin,Z. *et al.* (2016) Simultaneous dimension reduction and adjustment for confounding variation. *Proc. Natl. Acad. Sci. USA*, **113**, 14662–14667.

Lozupone,C. and Knight,R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.*, **71**, 8228–8235.

Lozupone,C.A. *et al*. (2007) Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.*, **73**, 1576–1585.

Pan,W. (2011) Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. *Genet. Epidemiol.*, **35**, 211–216.

Riquelme,E. *et al*. (2019) Tumor microbiome diversity and composition influence pancreatic cancer outcomes. *Cell*, **178**, 795–806.e12.

Wang,Y. *et al*. (2012) mvabund– an R package for model-based analysis of multivariate abundance data. *Methods Ecol. Evol.* **3**, 471–474.

Zapala,M.A. and Schork,N.J. (2006) Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proc. Natl. Acad. Sci. USA*, **103**, 19430–19435.