BIORANKINGS®
TECHNICAL REPORT SERIES

Dealing with High-Dimensional Data

Supplement #1: Example with Microbiome Data

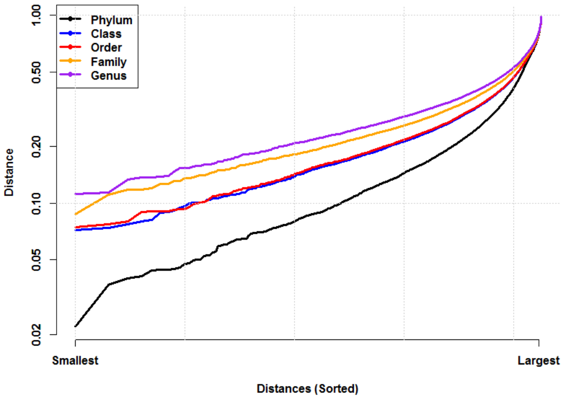
In ***BioRankings Technical Report #2, Dealing with High-Dimensional Data***, we discuss the weird behavior that occurs in pairwise distances when you have high-dimensional data, and how this behavior leads to incorrect statistical conclusions. (W Shannon, 2017) One of these behaviors is that as the number of dimensions increases, the distance between the two samples furthest apart gets closer to the distance between the two samples nearest each other. A result of this is that all pairwise distances become equal.

In **Technical Report 2** we showed through simulations that all pairwise distances become identical as the number of dimensions approaches infinity. This is a fact which can also be proven mathematically. In this **Supplement**, we demonstrate this theory with real microbiome data.

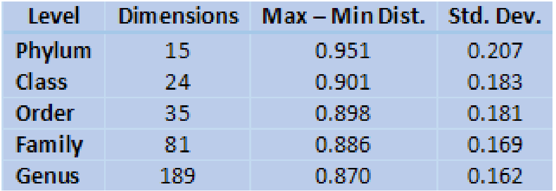
HMP Saliva Data

Taxa counts on 196 saliva samples from the HMP were used to create separate distance matrices at the Phylum, Class, Order, Family, and Genus taxa levels. Each matrix consisted of 17,000+ pairwise distances calculated using Bray-Curtis distance. The distances were sorted from smallest to largest. Pairwise between-subjects distances were calculated with dimensions ranging from 15 at the Phylum level to 189 at the Genus level.

As we expected, we found two patterns indicating the ***curse of dimensionality*** occurs for real data. First, as we moved from lower to higher dimensions, the difference between the maximum and minimum distances got smaller. Second, the pairwise distances overall became more similar, as reflected in the standard deviation.



The plot shows the 17,000+ pairwise distances (Y) sorted from smallest to largest (X) plotted for each taxa level. The pattern of the lines shows change in distances as we move from the lower dimension Phylum level (black line) to the higher dimension Genus level (purple line). Confirming what we expected to see, the smallest distances gets larger as the dimensions increase without a similar increase in the largest distances, and the lines are flatter moving from Phylum to Genus indicating the pairwise distances are becoming more similar with increasing dimensions.



Level	Dimensions	Max – Min Dist.	Std. Dev.
Phylum	15	0.951	0.207
Class	24	0.901	0.183
Order	35	0.898	0.181
Family	81	0.886	0.169
Genus	189	0.870	0.162

The table quantifies these patterns. First, the difference between the maximum and minimum distance becomes smaller with increasing dimensions and will continue to approach 0 as the dimensions increase. Second, the pairwise distances standard deviation decreases with increasing dimensions indicating that all the pairwise distances are approaching the same value.

Impact

This simple experiment with real microbiome data shows that the curse of dimensionality or large P small N problem impacts pairwise distances for even a relatively small number of dimensions. This should be considered when using distance-based methods. For example, PERMANOVA analyze distance matrices for hypothesis testing and power calculations. (Kelly et al., 2015) Since pairwise distances become more similar the results of this method will likely change for different dimensions. Cluster analysis is another distance based method whose results will be influenced by this phenomena. (W. Shannon, Culverhouse, & Duncan, 2003) As pairwise distances become closer in higher dimensions the samples become more uniformly spread out and cluster structure is destroyed.

BioRankings' Technical Report Series

BioRankings' mission is to help biomedical researchers move their technology from the lab to clinical applications using statistically valid analytical tools for efficient study designs, correct data analyses and conclusions, and rigorous and objective decision making for designing follow-up studies and eventual FDA approval.

To help achieve its mission, BioRankings publishes a Technical Report series focused on applying various statistical methods to real data analyses. Written for *understanding* by scientists and administrators, these reports will provide an intuitive understanding of the analyses leaving the statistical details to other publications.

For more information, contact BioRankings at 314-704-8725 or bill@biorankings.com.

References

Kelly, B. J., Gross, R., Bittinger, K., Sherrill-Mix, S., Lewis, J. D., Collman, R. G., . . . Li, H. (2015). Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA. Bioinformatics, 31(15), 2461-2468. doi:10.1093/bioinformatics/btv183

Shannon, W. (2017). Dealing with High Dimensional Data. Retrieved from http://biorankings.com/High-Dimensional_Data.pdf

Shannon, W., Culverhouse, R., & Duncan, J. (2003). Analyzing microarray data using cluster analysis. Pharmacogenomics, 4(1), 41-52. doi:10.1517/phgs.4.1.41.22581