



Generalized Canonical Correlation Analysis for Uncovering Microbiome Responses to Environmental Stress

Maria Hernandez Limon, Tannistha Mondal, Coralie Rousseau, Soutim Das, Cathy Pfister and Claire Donnat

In the Great Lakes, microbial communities play a crucial role in shaping ecosystems, influencing algal blooms and the food webs that support fisheries. Microbes can also form tightly coupled partnerships with algae (a “holobiont”), with important consequences for host health. Understanding interactions requires integrating and analyzing large multimodal datasets (e.g., sequencing, metabolites, host traits), yet many existing methods struggle with the high-dimensional and compositional nature of microbiome data and lack reliable uncertainty quantification. In this project, we develop new statistical methods for multimodal biological data integration, focusing on two core challenges:

1. *Compositional, high-dimensional multimodal integration via generalized CCA.* We build on Generalized Canonical Correlation Analysis (GCCA), a classical statistical method to jointly identify low-dimensional structure shared across multiple modalities. We adapt GCCA to high-dimensional settings by leveraging sparsity to improve inference whilst remaining computationally practical, with an emphasis on workflows that are compatible with compositional microbiome measurements (Aitchison geometry).

2. *Uncertainty quantification for CCA/GCCA.* To ensure reliable inference, we develop uncertainty quantification procedures for canonical correlations and canonical directions. We assess variable loadings by regressing each feature on held-out scores (and covariates when available), aggregating fold-wise t-tests using robust standard errors. We then apply empirical-Bayes shrinkage to obtain direction-specific effect sizes, confidence intervals, and local false sign rates with principled multiplicity control.

Applied to a longitudinal microbiome dataset from the Great Lakes, our approach links microbial community structure across distinct size fractions to environmental drivers. We are finalizing an open-source (GCCA) software package, with explicit support for compositional data, and are deploying these tools across both the Great Lakes and kelp-associated microbial communities to identify microbial communities and taxa that support host resilience to environmental stress.

Reference: <https://github.com/donnate/SGCAR/tree/r-package>. This research was supported in part by grants from the NSF (DMS-2235451) and Simons Foundation (MP-TMPS-00005320).

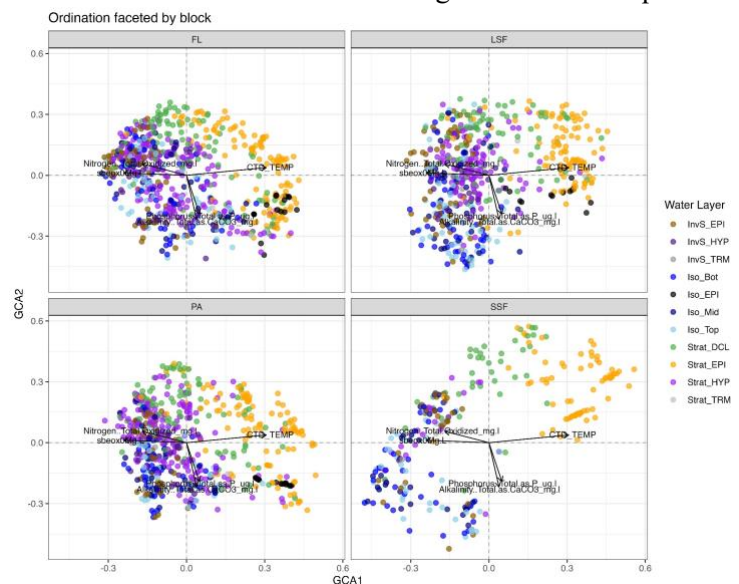


Figure 1. Visualization of the GCA-aligned microbiome compositions, faceted by size fraction. GCA reveals strong associations between water column and composition.