Hypothetical Example: Ordering land/water samples according to unknown environmental gradient (gradient hypothesized to be salinity) 

X marks sample site: bacteria abundance data obtained via 16s rRNA/pyrosequencing

The spatial ordering of samples are: PA, Del River, NJW, NJE, Bay, LBI, Atlantic

We wish to order these samples by the underlying gradient, which should be causing the species abundances to vary.

Expected output (if salinity is the gradient): PA, NJW, NJE, LBI, Del River, Bay, Atlantic

Increasing Salinity

Empirical Example: Bacteria have been sampled along a pH gradient. Can we recover the sample order when the gradient is decreasing?

Controlled site so that pH is the only major influence

\[ X_1 \leq X_2 \leq X_3 \leq \ldots \leq X_{22} \]

\[ pH = 4.1 \leq 8.3 \]

X marks sample site: bacteria abundance data obtained via 16s rRNA/pyrosequencing

The output of the analysis should produce samples ordered 1 to 22 or 22 to 1

A typical approach to the problem: PCA Analysis

1. Standardize community data matrix (subtract mean, divide by std. dev.)
2. Compute symmetric correlation matrix [\( S \) x \( S \)]
3. Find orthogonal vectors of maximum variance (Eigen-analysis)
4. Keep two eigenvectors corresponding with the two largest eigenvalues
5. Plot PCAx vs. PCAy;

Considerations:

1. Arch Effect: Characteristic curve is an artifact of the method - obscures interpretation of PCAx
2. Cannot resolve or detect multiple gradients due to arch effect
3. Effectively, most gradient influence is compacted in PCAx (highest variance) so we cannot say that PCAy is the gradient.
4. Assumes that the species are monotonically related to one another and the gradients.

Novel Approach to the Problem:

Fitness Function

Given a permutation of samples \( x \), test the null hypothesis that the abundance of each species is randomly distributed across the samples.

**Wald-Wolfowitz Runs Test, \( w(x) \)**

\[ y = \sum_{i=1}^{m} h(x_i) \left( \{ x \in [0, m] \} \text{ and } \{ y \in [0, 1] \} \right) \]

Expected output (if salinity is the gradient): PA, NJW, NJE, LBI, Del River, Bay, Atlantic

Particle Swarm Optimization

Inspiried by the behavior of insects in a population.

We model each insect as a particle representing one of our samples, exist in the solution space.

Each particle has a position and a velocity. With each iteration of the PSO algorithm, a particle moves in the solution space based on its performance on the fitness function and that of the entire swarm.

The benefit of this behavior is that in many cases the optimization avoids local minima.

**Performance: Proposed Algorithm**

*This work was supported in part by National Science Foundation award #0849527*