ReadMe for localPCA Folder

Notation:

- $\{X_i : i = 1, \ldots, N\}$ = input data set
- $p(X_i, d)$ = projection of X_i onto first d basis vectors resulting from PCA (basis vectors are in order of most to least variance)
- μ_{global} = global mean, i.e. mean of entire data set
- $\mu_{\text{local},i}$ = local mean, i.e. mean of neighborhood with center X_i

`local_pca.m`

Here, we randomly choose n ($\leq N$) center points $\{X_i : i = 1, \ldots, n\}$. For each center point, we label its k neighbors as $\{X_{i,j} : j = 1, \ldots, k\}$, where $X_{i,1} = X_i$ and neighbors are labeled in order of ascending distance from the center point. Currently, we compute the output error fraction as:

$$T = \frac{1}{n} \sum_{i=1}^{n} \frac{\sum_{j=1}^{k} \|X_{i,j} - p(X_{i,j}, d)\|^2}{\sum_{j=1}^{k} \|X_{i,j} - \mu_{\text{global}}\|^2}.$$

`local_pca_2k.m`

Here, we randomly choose n ($\leq N$) center points $\{X_i : i = 1, \ldots, n\}$. For each center point, we label its $2k$ neighbors as $\{X_{i,j} : j = 1, \ldots, 2k\}$, where $X_{i,1} = X_i$ and neighbors are labeled in order of ascending distance from the center point. We perform PCA on half of the neighbors, i.e. $\{X_{i,2j+1} : j = 0, \ldots, k-1\}$. Then, we compute the output error fraction on the remaining half of the neighborhood:

$$T = \frac{1}{n} \sum_{i=1}^{n} \frac{\sum_{j=1}^{k} \|X_{i,2j} - p(X_{i,2j}, d)\|^2}{\sum_{j=1}^{k} \|X_{i,2j} - \mu_{\text{global}}\|^2}.$$

Note: We were considering alternative formulations of the local PCA algorithm in hopes of being able to see more of a distinction between the behavior of linear and nonlinear data. Some of these ideas included:

- Forcing PCA to go through the center point X_i rather than through the local mean $\mu_{\text{local},i}$
- Replacing μ_{global} with $\mu_{\text{local},i}$ in both of the above formulas for T
- Looking at how PCA coefficients, i.e. the orientation of the plane, changes with neighborhood size
- Looking at T vs. k (instead of T vs. d)