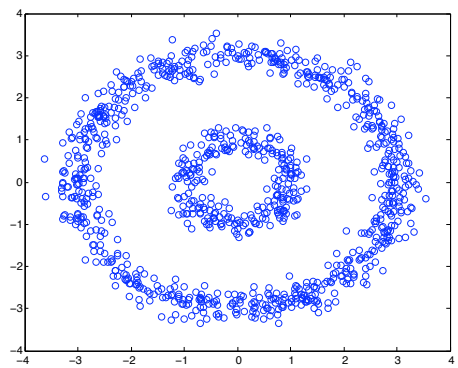
CSE 291F. Assignment 3

Out: *Tue May 25*
Due: *Thu Jun 10*

3.1 Spectral clustering versus k-means

Download the *rings* data set for this problem from the course web site. The data is stored in MATLAB format as a matrix X with D rows and N columns, where $D = 2$ is the input dimensionality and $N = 1000$ is the number of inputs. Submit your source code as part of your solutions.



(a) **Spectral clustering**

Compute a binary partition of the data using spectral clustering with affinities $K_{ij} = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma^2}$. For what values of the kernel width σ does the method separate the two rings? Visualize your results in the plane.

(b) **k-means clustering**

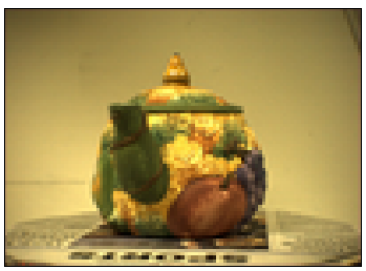
Use your best results from spectral clustering to initialize a k-means algorithm with $k = 2$. Does k-means clustering preserve or modify your previous solution? For better or for worse?

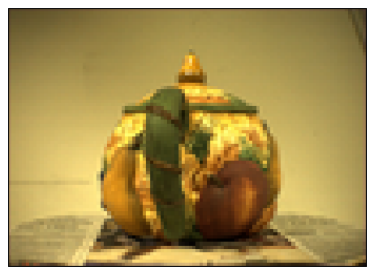
3.2 Dimensionality reduction of images

Download the *teapot* data set for this problem from the course web site. The images are stored in MATLAB format as a matrix X with D rows and N columns, where $D = 23028$ is the number of pixels per image and $N = 400$ is the number of images. The ith image in the data set can be displayed using the command:

$$\texttt{imagesc}(\texttt{reshape}(\texttt{X}(:, \texttt{i}), 76, 101, 3))$$

In this problem you will compare the results from linear and nonlinear methods for dimensionality reduction. For the latter, you may download and make use of publicly available code for any of the algorithms we discussed in class.





(a) **Subspace methods**

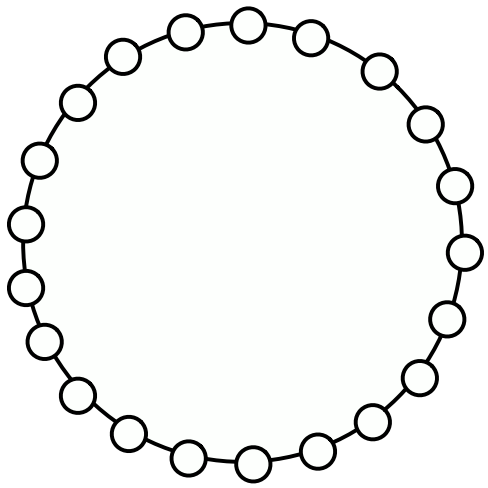
Center the data by subtracting out the mean image, then compute the Gram matrix of the centered images. Plot the eigenvalues of the Gram matrix in descending order. How many leading dimensions of the data are needed to account for 95% of the data's total variance? Submit your source code as part of your solutions.

(b) **Manifold learning**

Use a manifold learning algorithm *of your choice* to compute a two dimensional embedding of these images. By visualizing your results in the plane, compare the two dimensional embeddings obtained from subspace methods versus manifold learning.

3.3 Sparse matrices on ring graphs

Consider a ring graph with N nodes and uniform weights on its edges. In particular, let $W_{ij} = \frac{1}{2}$ if $|j-i| = 1$ or if $|j-i| = N-1$; otherwise $W_{ij} = 0$.



(a) **Graph Laplacian**

Show that the normalized graph Laplacian is a circulant matrix, and use this property to compute all of its eigenvalues and eigenvectors.

(b) **Locally linear embedding (LLE)**

Consider the matrix $(I-W)^\top(I-W)$ diagonalized by LLE. How are its eigenvalues and eigenvectors related to those of the normalized graph Laplacian?

3.4 Isomap on a ring

Consider N data points that are uniformly sampled along the unit circle in the plane. In this problem, you will study the behavior of the Isomap algorithm in the continuum limit $N \to \infty$. In this limit of large sample size, it is possible to compute an analytical solution. This solution reveals how Isomap breaks down when its convexity assumption is violated.

(a) **Inner products**

Applied to a finite data set, Isomap estimates the geodesic distances $\mathcal{D}_{ij}$ between the ith and jth inputs, then constructs the Gram matrix with elements:

$$G_{ij} = \frac{1}{2}\left[\frac{1}{N}\sum_k \left(\mathcal{D}_{ik}^2 + \mathcal{D}_{kj}^2\right) \;-\; \frac{1}{N^2}\sum_{k\ell} \mathcal{D}_{k\ell}^2 \;-\; \mathcal{D}_{ij}^2\right].$$

In the continuum limit $N \to \infty$, a data point exists at every angle $\theta \in [0, 2\pi]$ on the unit circle. The geodesic distance between two points at angles θ and θ' is given by:

$$\mathcal{D}(\theta, \theta') = \min(|\theta - \theta'|, 2\pi - |\theta - \theta'|).$$

In this limit, the Gram matrix normally constructed by Isomap is replaced by a kernel function $g(\theta, \theta')$. The kernel function can be computed from:

$$g(\theta, \theta') = \frac{1}{2}\left[\int_0^{2\pi} \frac{d\phi}{2\pi}\left(\mathcal{D}^2(\theta, \phi) + \mathcal{D}^2(\phi, \theta')\right) \;-\; \int_0^{2\pi}\frac{d\phi}{2\pi}\int_0^{2\pi}\frac{d\psi}{2\pi}\,\mathcal{D}^2(\phi, \psi) \;-\; \mathcal{D}^2(\theta, \theta')\right].$$

Evaluate these integrals to derive an expression for the kernel function $g(\theta, \theta')$. In particular, show that in the continuum limit:

$$g(\theta, \theta') \;=\; \frac{\pi^2}{6} - \frac{1}{2}\mathcal{D}^2(\theta, \theta').$$

Plot the kernel function $g(\theta, \theta')$ as a value of the difference in angle $|\theta - \theta'|$. On the same figure, plot $k(\theta, \theta') = \cos(\theta - \theta')$, which measures the true inner products of points on the unit circle. How are these plots different?

(b) **Eigenvalues**

Applied to a finite data set, Isomap computes the eigenvalues and eigenvectors of the Gram matrix estimated from the first equation in part (a). In the continuum limit $N \to \infty$, Isomap's output is encoded by the eigenvalues λ_n and eigenfunctions $u_n(\theta)$ of the kernel function $g(\theta, \theta')$. These eigenvalues and eigenfunctions satisfy:

$$\int_0^{2\pi} d\phi\; g(\theta, \phi)\, u_n(\phi) \;=\; \lambda_n u_n(\theta).$$

Show that $\sin(r\theta)$ and $\cos(r\theta)$ are eigenfunctions of the kernel function $g(\theta, \theta')$ for integer-valued $r \in \{0, 1, 2, \ldots, \infty\}$. (Ignore the degenerate case $\sin(r\theta)$ for $r = 0$.) What are the eigenvalues of these eigenfunctions? How many of them are non-zero? Is there a gap in the eigenvalue spectrum that reveals the intrinsic dimensionality of the ring?