STAT 185 - Problem Set #5

NAME

November 5, 2019

Instructions

This assignment covers Multidimensional Scaling and nonlinear manifolds through a mixture of theory and coding. You will need the file *pset5.Rdata* available on Canvas. The commands for loading the data assume this file, *pset5.Rmd*, and the Rdata file are in the same folder.

You can use the command *cmdscale(d, k , eig = FALSE)* to compute the solution to the classical scaling problem. Here d is a matrix of distances, k is the dimension of the desired configuration, and the option *eig=TRUE* may be added so that eigenvalues are also reported.

For metric MDS, use the command *sammon(d, k, trace = FALSE)* for metric MDS using Sammon's Stress. Here d is a matrix of distances, k is the dimension of the desired configuration, and the option *trace = FALSE* stops reporting of the stress over iterations of the algorithm.

Use the command *isoMDS(d, k, trace = TRUE)* for nonmetric MDS using the Kruskal Stress-1. As above, d is a matrix of distances, k is the dimension of the desired configuration, and the option *trace = FALSE* stops reporting of the stress over iterations of the algorithm.

Given a data matrix, all pairwise distances can be calculated using the command *dist(x)* where x is the data matrix. The default method is the Euclidean distance. All of the above MDS commands require a matrix of distances or variable of type *dist* given by the output of the *dist()* function.

Complete the following questions and submit your solutions as a *pdf* via Canvas by 11:59 PM on Tuesday November 5th.

Problems

1. In this problem, we are going to derive the update rule for minimizing Sammon Stress. Given dissimilarities $\delta_{i,j}$ for $1 \le i < j \le N$, recall the metric MDS Sammon Stress formula,

$$Stress(W, \vec{y}_1, \dots, \vec{y_N}) = S = \sqrt{\sum_{i,j} W_{ij}(d_{ij} - \delta_{ij})^2}.$$

where d_{ij} is the distance from $\vec{y}_i$ to $\vec{y}_j$ in $\mathbb{R}^k$ under the Euclidean norm ($d_{ij}^2 = \sum_k (y_{ik} - y_{jk})^2$) and weights $W_{ij} = \delta_{ij}^{-1}\left[\sum_{i<j} \delta_{ij}\right]^{-1}$ depending only on the dissimilarities. Sammon developed a method which uses steepest descent, so that if $y_{ij}^{(m)}$ is the mth iteration in minimising S^2, then

$$y_{ij}^{(m+1)} = y_{ij}^{(m)} - MF\left(\frac{\partial(S^2)}{\partial y_{ij}}\right) / \left|\frac{\partial^2(S^2)}{\partial y_{ij}^2}\right|$$

where MF is a *magic factor* to optimize convergence of the algorithm.

   a. Find the partial derivative of S^2 with respect to y_{ij}, the jth coordinate of the ith point in the configuration.

   b. Find the second partial derivative of S^2 with respect to y_{ij}, i.e. $\frac{\partial(S^2)}{\partial y_{ij}^2}$.

2. In some cases, similarities are provided rather than dissimilarities or distances. A matrix $C \in \mathbb{R}^{N\times N}$ is a similarity matrix if it is symmetric and

$$c_{rs} \leq c_{rr}, \qquad \text{for all } r, s.$$

The standard transformation from a similarity matrix C to a distance matrix D is defined by

$$d_{rs} = (c_{rr} - 2c_{rs} + c_{ss})^{1/2}.$$

For all subproblems, assume that C is a positive semidefinite similarity matrix. (Note, a matrix, $A \in \mathbb{R}^{N\times N}$, is positive semidefinite if $\vec{v}^T A\vec{v} \geq 0$ for all vectors $\vec{v} \in \mathbb{R}^N$.)

a. Show that $c_{ii} + c_{jj} - 2c_{ij} \geq 0$.

b. Show that the distances defined by $d_{rs}^2 = c_{rr} + c_{ss} - 2c_{rs}$ satisfy the triangle inequality

$$d_{rs} + d_{st} \geq d_{rt}, \text{ for all } r, s, t.$$

c. If C is a positive semidefinite similarity matrix, show that D is Euclidean with centered inner product matrix $B = HCH$ where H is the centering matrix $H = I_N - \frac{1}{N}1_N 1_N^T$ and 1_N is the vectors in $\mathbb{R}^N$ comprised of all ones.

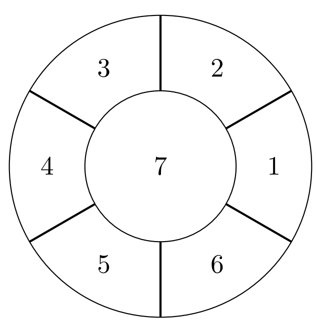
3. Consider the color-stimuli experiment discussed in class and in the Izenman text. The *similarity* ratings are given in the variable *problem2*. The column names correspond to the wavelength of the colors.

a. Compute the dissimilarity matrix.

b. Carry out classical scaling of the data. Generate a plot of all eigenvalues and the configuration in 2-dimensions.

c. Carry out nonmetric scaling of the data. Generate a plot of the Kruskal Stress as a function of dimension and show the configuration in 2-dimensions. (See Figure 13.3 of the Izenman text for reference).

4. Suppose that $1, \ldots, 7$ are regions enclosed by unbroken lines in a country shown in the arrangement below.



a. Suppose the distance matrix is constructed by counting the minimum number of boundaries crossed to pass from region i to region j. Give the distance matrix, D.

b. Show that the distances constructed this way obey the triangle equality.

c. Find the eigenvalues in the classical scaling. Is D Euclidean?

d. Plot coordinates given by the solution to the classical scaling problem in 2-dimensions. Is the original map, from the figure above, reconstructed?

5. Let $D \in \mathbb{R}^{N\times N}$ be Euclidean distance matrix with configuration $\mathbf{X} = [\vec{x}_1, \ldots, \vec{x}_N]^T$ given by the p-dimensional principal component scores. Suppose we wish to add an additional point to the configuration usings distances $d_{r,N+1}$ for $r = 1, \ldots, N$ which are also Euclidean, allowing for a $(p+1)$-dimensional

configuration. If the first N points in the $(p+1)$-dimensional configuration are $\vec{x}_r = (x_{r1}, \ldots, x_{rp}, 0)^T$ for $r = 1, \ldots, N$, then show that the $(N+1)$th point is given by $\vec{x}_{N+1} = (\vec{x}^T, y)^T$ where

$$\vec{x} = \frac{1}{2}\mathbf{\Lambda^{-1}}\mathbf{X}^T\vec{f}, \qquad \vec{f} = (f_1, \ldots, f_N)^T, \qquad f_r = b_{rr} - d^2_{r,N+1}$$

and

$$y^2 = \frac{1}{N}\sum_{r=1}^{N} d^2_{r,N+1} - \frac{1}{N}\sum_{r=1}^{N} b_{rr} - \|\vec{x}\|^2.$$

Hence, $\vec{x}$ is uniquely determined but y is determined in absolute value but not in sign.

6. The variables *problem6* contains points on a helix embedded in $\mathbb{R}^{50}$.

   a. Compute the distance matrix under the Euclidean norm and plot the eigenvalues of the classical scaling as a function of dimension. Briefly, discuss the figure and its implications about the dimensionality of the data.

   b. Show the 1, 2, and 3 dimensional configurations in the classical scaling and discuss the differences. Plots in 3D can be generated using the command *scatterplot3d*.