

BINF 733 CSI739 – Spring 2004 Weller/Solka
Dimensionality Reduction HW - 2

Obtain the Golub data set and class files from our 2003 class website <http://www.binf.gmu.edu/%7Ejsolka/s2003/csi739/csi73903.htm>. Please try to have the problems completed by 3/18/04. Email me your solutions, jsolka@gmu.edu, as a pdf file if possible.

#1 (20 pts) View the data as 72 observations in 7129 dimensions. Plot the AML, ALL T-cell and ALL B-cell observations in the first two components obtained via multi-dimensional scaling in R. Use a different symbol and color for the AML, ALL T-cell and ALL B-cell observations. Provide code and 3 two-dimensional scatter plots (one obtained using `cmdscale`, one obtained using `isoMDS`, and one obtained using `sammon`).

#2 (20 pts) View the data as 72 observations in 7129 dimensions. Plot the AML, ALL T-cell and ALL B-cell observations in the first two components obtained via an ISOMAP projection in MATLAB. Use a different symbol and color for the AML, ALL T-cell and ALL B-cell observations. I think that you should set an upper bound on the ISOMAP dimension of say 150. Provide code to produce a two-dimensional scatter plot and the ISOMAP scree plot. What dimensionality value does ISOMAP suggest for the AML ALL data? Does ISOMAP suggest that the data resides in one contiguous component?

#3 (20 pts) One way to ascertain the appropriate dimensionality in MDS is make a scree plot of returned stress vs. the number of dimensions. Write an R function that will perform this exercise for the GOLUB data using the R `sammon` function. I would suggest that you might start at say $k = 150$ and plot stress as a function of k for all of the integer values between $k = 150$ and $k = 2$. Where is the elbow in this curve?

#4 (20 pts) Parallel coordinates is one way to examine the structure of moderately high dimensional data. Check out the help in R on their version of the parallel coordinate command using `?parcoord`. I am not

sure what dimension value will be returned by problem #3. If this value is less than or equal to twenty then plot the ALL T cell observations, the AML B-cell observations and the AML observations using the parcoord command in different colors. If this value is larger than 20 then do this exercise based on a projection to 20 dimensions.

#5 – Repeat your ISOMAP analysis of Problem # 2. This time ascertain what the elbow in the curve, intrinsic dimensionality, is for the ALL T cell, the ALL B cell and the AML observations. This will require you to run each of these datasets in the ISOMAP procedure separately. What are the 3 intrinsic dimensionalities reported by ISOMAP?