Math 585 HW 7 Spring 2024 Due Friday, March 29
Quiz 7 on Wed. March 27, covers HW6 and HW7. 2 pages, problems A)-D)

For the following R problems perform the perform the *source("J:/mpack.txt")* and *source("J:/mrobdata.txt")* commands as described in homework 3. Also copy and paste commands from (http://parker.ad.siu.edu/Olive/mrsashw.txt) for the relevant problem into R.

A), 6.11 The R data set USArrests contains statistics, in arrests per 100,000 residents, for assault, murder, and rape in each of the 50 US states in 1973. The fourth variable, UrbanPop, is the percent urban population in each state. For PCA, the R `summary` command can be used to get proportion of variance explained and cumulative proportion of variance explained, similar to *SAS* output.

a) Use the R *commands* for this part to get the classical and robust PCA summaries where $\boldsymbol{S}$ or $\boldsymbol{S}_U$ is used. Paste the summaries into *Word*.

i) Are the summaries similar?

ii) Using the 0.9 threshold, how many principal components are needed?

a) Use the R *commands* for this part to get the classical and robust PCA summaries where $\boldsymbol{R}$ or $\boldsymbol{R}_U$ is used. Paste the summaries into *Word*.

i) Are the summaries similar?

ii) using the 0.9 threshold, how many principal components are needed?

B), 8.6 Wisseman, Hopke and Schindler-Kaudelka (1987) pottery data has 36 pottery shards of Roman earthware produced between second century B.C. and fourth century A.D. Often the pottery was stamped by the manufacturer. A chemical analysis was done for 20 chemicals (variables), and 28 cases were classified as Arrentine (group 1) or nonArrentine (group 2), while 8 cases were of questionable origin. So the training data has $n = 28$ and $p = 20$.

a) Copy and paste the R commands for this part into R to make the data set.

b) Because of the small sample size, LDA should be used instead of QDA, as in the handout. Nonetheless, variable selection using QDA will be done. Copy and paste the R commands for this part into R. The first 9 variables result in no misclassification errors.

c) Now use commands like those shown in this section to delete variables whose deletion does not result in a classification error. Should get four variables are needed for perfect classification. What are they (eg X1, X2, X3 and X4)?

C), 8.7 Variable selection for LDA used the pottery data described in Problem 8.6, and suggested that variables X6, X11, X14, and X18 are good. See Example 8.6. Use the R commands for this problem to get the apparent error rate AER.

D), 8.8 The distance discriminant rule is attractive theoretically as a maximum likelihood discriminant rule, but the distance rule does not work well for groups that have similar means. The `ddiscr` rule is a modification of the distance rule, and the `ddiscr2` rule tries to use the maximum likelihood rule where the $\hat{f}_j$ are estimated with a kernel density estimator.

The R code for this problem generates $N_2(\mathbf{0}, \boldsymbol{I})$ data where group 1 consists of the half set of cases closes to $\mathbf{0}$ in Mahalanobis distance (an ellipse about the origin), and group 2 consists of the remaining cases (the covering ellipse with inner ellipse removed).

a) Copy and paste the commands to make the data.

b) The commands for this part give the error rate for the ddiscr method that uses $\boldsymbol{x}$ as the two predictors. Put this output in *Word*.

c) The commands for this part give the error rate for the ddiscr method that uses the distances based on group 1 applied to all of the cases as the predictor. Put this output in *Word*.

d) The commands for this part give the error rate for the ddiscr2 method that uses $\boldsymbol{x}$ as the two predictors. Put this output in *Word*.

e) The commands for this part give the error rate for the ddiscr2 method that uses the distances based on group 1 applied to all of the cases as the predictor. Put this output in *Word*.

f) The commands for this part get the error rate for LDA using $\boldsymbol{x}$ as the two predictors.

g) The commands for this part get the error rate for QDA using $\boldsymbol{x}$ as the two predictors.

h) Which method worked the best?