`Math 585    HW 8 Spring 2024        Due Friday, April 5`
`Quiz 8 on Wed. April. 3, covers HW8. 2 pages, Problems A)-E)`

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), 4.9 The R function `cov.mcd` is a Fake-MCD estimator. If `cov.mcd` computed the minimum covariance determinant estimator, then the log determinant of the dispersion matrix would be a minimum and would not change when the rows of the data matrix are permuted. The *R commands* for this problem permute the rows of the Gladstone (1905-6) data matrix seven times. The log determinant is given for each of the resulting `cov.mcd` estimators.

a) Paste the output into *Word*.

b) How many distinct values of the log determinant were produced? (Only one if the MCD estimator is being computed.)

B), 6.12 For PCA, a *biplot* is a plot of the first principal component versus the second principal component. The plotted points are $\hat{\boldsymbol{e}}_j^T \boldsymbol{x}_i$ for $j = 1, 2$ where the classical biplot uses $i = 1, ..., n$ and the robust plot uses cases in the RMVN set U. Let $\hat{\boldsymbol{e}}_j = (\hat{e}_{1j}, \hat{e}_{2j}, ..., \hat{e}_{pj})^T$. Then $\hat{e}_{kj}$ is called the *loading* of the kth variable on the jth principal component. An arrow with the kth variable name is the vector from the origin $(0, 0)^T$ to the loadings $(\hat{e}_{k1}, \hat{e}_{k2})^T$. So if the arrow is in the first quadrant, both loadings are positive, etc. If the arrow is long to the right but short down, then the loading with the first principal component is large and positive while the loading with the second principal component is small and negative.

The Buxton (1920) data has a cluster of 5 massive outliers. The first classical principal component tends to go right through a cluster of large outliers.

a) These R commands make the classical scree plot and biplot. Paste the plots into *Word*.

b) These R commands make the robust scree plot and biplot. Paste the plots into *Word*.

c) From the classical scree plot, how many principal components are needed? From the robust scree plot, how many principal components are needed?

d) The four variables used were *len, nasal, bigonal*, and *cephalic* . From the classical biplot, which variable had the 5 massive outliers.

e) From the robust biplot, which two variables loaded highest with the first principal component?

C), 9.5 Conjecture:

$$\sqrt{n}(T_{RMVN} - \boldsymbol{\mu}) \xrightarrow{D} N_p(\mathbf{0}, \tau_p \boldsymbol{\Sigma})$$

for a wide variety of elliptically contoured distributions where τ_p depends on both p and the underlying distribution. The following "test" is based on a conjecture, and should be used as an outlier diagnostic rather than for inference. The ad hoc "test" that rejects H_0 if

$$T_R^2/f_{n,p} = n(T_{RMVN} - \boldsymbol{\mu}_0)^T \hat{\boldsymbol{C}}_{RMVN}^{-1} (T_{RMVN} - \boldsymbol{\mu}_0)/f_{n,p} > \frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}$$

where $f_{n,p} = 1.04 + 0.12/p + (40 + p)/n$. The simulations use $n = 150$ and $p = 10$.

a) The *R* commands for this part use simulated data is $\boldsymbol{x}_i \sim N_p(\mathbf{0}, diag(1, 2, ..., p))$ where $H_0 : \boldsymbol{\mu} = \mathbf{0}$ is being tested with 5000 runs at a nominal level of 0.05. So H_0 is true, and hcv and rhcv are the proportion of rejections by the T_H^2 test and by the ad hoc robust test. Want hcv and rhcv near 0.05. THIS SIMULATION WILL TAKE ABOUT 5 MINUTES. Record hcv and rhcv. Were hcv and rhcv near 0.05?

b) The *R* commands for this part use simulated data is $\boldsymbol{x}_i \sim N_p(\delta\mathbf{1}, diag(1, 2, ..., p))$ where $H_0 : \boldsymbol{\mu} = \mathbf{0}$ is being tested with 5000 runs at a nominal level of 0.05. In the simulation, $\delta = 0.2$, so H_0 is false, and hcv and rhcv are the proportion of rejections by the T_H^2 test and by the ad hoc robust test. Want hcv and rhcv near 1 so that the power is high. Paste the output into *Word*. THIS SIMULATION WILL TAKE ABOUT 5 MINUTES. Record hcv and rhcv. Were hcv and rhcv near 1?

D), 9.1 The Wechsler Adult Intelligence Scale scores of $n = 101$ subjects aged 60 to 64 were recorded, giving a verbal score (X_1) and performance score (X_2) for each subject. Suppose $\boldsymbol{\mu}_0 = (60, 50)^T$ and $T_C^2 = 357.43$. Perform the one sample Hotelling's T^2 test.

E), 9.2 The levels of free fatty acid (FFA) in the blood were measured in $n = 15$ hypnotized normal volunteers who had been asked to experience fear, depression, and anger effects while in the hypnotic state. The mean FFA changes were $\overline{x}_1 = 2.669$, $\overline{x}_2 = 2.178$, and $\overline{x}_3 = 2.558$. Let $\mu_F = \mu + \tau_1$, $\mu_D = \mu + \tau_2$ and $\mu_A = \mu + \tau_3$. Want to know if the mean stress FFA changes were equal. So test whether $\mu_F = \mu_D = \mu_F$ if $T_R^2 = 2.68$.