THE UNIVERSITY OF MINNESOTA

Statistics 5401/8401 September 16, 2005

Assignment Sheet No. 2

**Reading**

Week of September 12-16: J&W Chapter 3
Week of September 19-23: J&W , Chapter 4 (additional handouts will be posted)

**Written Assignment** (due in class, Friday, September 23):
Use the data from J&W Table 1.6 (p. 42, dataset `T01_06` in file `JWData5.txt`). You can read it by

```
Cmd> jw1_6 <- read("","t01_06") # or matread()
```

and seach for `JWData5.txt` using the file navigation dialog box.

Or, if you set `DATAFILE` by

```
Cmd> DATAFILE <- getfile() # find JWData5.txt in dialog box
```

After `getfilename()` This brings up a file navigation dialog box, find `JWData5.txt` and select it. Then you can read the data by

```
Cmd> jw1_6 <- getdata(t01_06)
```

You can get a list of all the datasets in `JWData5.txt` and a discription of how they are named by

```
Cmd> read("") # find JWData5.txt
```

The data set contains data on two groups of subjects, those without multiple sclerosis (non-MS group) and those with multiple sclerosis (MS group). Let $\bar{\boldsymbol{x}}_1$, $\bar{\boldsymbol{x}}_2$, $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ be the sample means and unbiased (divisor $n_i - 1$) sample variance matrices of the non-MS and MS groups, omitting, of course, column 1, the group number.

You should first create 69 by 5 and 29 by 5 matrices `nonms` and `ms`, respectively by

```
Cmd> group <- jw1_6[,1]
Cmd> nonms <- jw1_6[group == 1,-1]
Cmd> ms <- jw1_6[group == 2,-1]
```

**Note**: The first subscript, `group == 1`, is a `LOGICAL` subscript with values True for all cases for which `group` (from column 1 of `x`) is 1 and False otherwise; when a `LOGICAL` subscript is the first subscript of a matrix, it selects all the rows corresponding to True. Thus this selects only the cases for which column 1 is 1. The second subscript (`-1`) means omit column 1.

STAT 5401/8401 Assignment Sheet No. 2

1 Make a **faces plot** with the first 10 cases of each group in a single plot. Make another with the variables in reverse order ($x_5$, $x_4$, $x_3$, $x_2$, $x_1$). Comment on how well the faces could help you group together similar cases, and on the difference between the two plots.

You will need both the new macro file `mvgraphics.mac` and the corrected macro fiole `graphics.mac`, both of which are available on the web. As illustrated before, you will need to do something like the following:

```
Cmd> addmacrofile("")  # find mvgraphics.mac and select it
Cmd> addmacrofile("")  # find the new graphics.mac
```

Note: The second `addmacrofile()` is not needed on the School of Statistics work stations as the revised `graphics.mac` is already installed there.

One way get a data matrix containing the cases whose faces are to be drawn is the following:

```
Cmd> selected <- vconcat(nonms[run(10),], ms[run(10),])
Cmd> faces(selected,\
    title:"Chernoff faces for multiple sclerosis data",\
    xlab:"1st 10 are non-MS and last 10 are MS")
```

To make the plot with the variables in reverse order, do either of the following:

```
Cmd> faces(selected[,run(5,1)]) #run(5,1) = vector(5,4,3,2,1)
Cmd> faces(selected, whichvars:run(5,1))
```

I left out a title and x axis label for brevity. You should not do so.

2. Make an **Andrews plot** of the entire data with the variables in the original order and in the reverse order and comment on the differences and whether either plot is informative. You can make the plots (omitting title and label) by

```
Cmd> andrewsplot(jw1_6[,vector(2,3,4,5,6)],jw1_6[,1])
Cmd> andrewsplot(jw1_6[,vector(6,5,4,3,2)],jw1_6[,1])
```

3. Assess the marginal normality of variables $x_2$, $x_3$, and $x_4$ for the both the MS and non-MS groups. You can make a normal scores plot of variable 2 of `ms`, say, by

```
Cmd> plot(rankits(ms[,2]),Scler_2:ms[,2], xlab:"Normal scores",\
     ylab:"S1L + S1R", \
     title:"Normal scores plot of S1L + S1R of Sclerosis data")
```

Comment on the apparent normality or lack of normality.

**Note**: When `x` is a (column) vector `rankits(x)` computes normal scores for each case element of `x`, maintaining the ordering of the data. Note the use of keywords `xlab`, `ylab` and `title` to provide axis labels and a heading for the graph.

3. Use `distcomp()` to compute the generalized distances $d_{jk}{}^2 = (\boldsymbol{x}_{jk} - \bar{\boldsymbol{x}}_k)' S_k^{-1} (\boldsymbol{x}_{jk} - \bar{\boldsymbol{x}}_k)$, $j = 1,\ldots,n$ for groups $k = 1,2$.

Make a $\chi^2$ Q-Q plot for the full data from each group to assess multivariate normality. This can be done as follows if the data are in matrix `y`:

```
Cmd> d <- distcomp(y)
WARNING: searching for unrecognized macro distcomp near d <-
distcomp(

Cmd> p <- ncols(y); n <- nrows(y)

Cmd> probs <- (run(n)-.5)/n #equally spaced from .5/n to 1-.5/n

Cmd> x <- invchi(probs,p) # chi-squared_p probability points

Cmd> plot(x,sort(d),title:paste("Q-Q plot with df =",p),\
         xlab:"Chi-squared quantiles",ylab:"Distances")
```

**Note**: `distcomp()` is a macro in file `Mulvar.mac` which is installed with MacAnova. Using such a macro causes it to be read from the file. The `WARNING` line basically means MacAnova is about to search in the file for the macro.

4. Compute a pooled variance matrix using both MS and non-MS data

$$\boldsymbol{S} = (n_1+n_2-2)^{-1}\{(n_1-1)\boldsymbol{S}_1 + (\mathrm{n}_2-1)\boldsymbol{S}_2\}$$

and compute the standardized squared distance between $\bar{\boldsymbol{x}}_1$ and $\bar{\boldsymbol{x}}_2$

$$\mathrm{T}^2 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\hat{V}[\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2]^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\{(\frac{1}{n_1} + \frac{1}{n_2})\boldsymbol{S}\}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$$

<u>Remark</u>: Provided $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, $\hat{V}[\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2] = (1/n_1 + 1/n_2)\,\boldsymbol{S}$ estimates

$$(1/n_1 + 1/n_2)\boldsymbol{\Sigma} = V[\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2].$$