

Assignment Sheet No. 1

Reading (for week of September 7-9):

Review Johnson and Wichern (J&W) Chapters 1, 2 (Generally proofs and most of the mathematical definitions involving integration and differentiation are not important)

Familiarize yourself with MacAnova.

If you are new to MacAnova, start with *An Introduction to MacAnova*, and work through the examples at a computer. On the class web page, <http://www.stat.umn/~kb/classes/5401> there is a link to a page to download an Acrobat PDF version of this document (click on *Introduction to MacAnova*). This is not up-to-date and some screen shots don't exactly match what you will see.

MacAnova Users' Guide is a comprehensive manual for Version 4.07 and does not include features added in later versions. It is available as a set of PDF files on the World Wide Web. This is a link to these files on the class web page (click on User's Guide for MacAnova 4.07). I don't recommend you download the whole thing, although you may want to at some point.

You will need Acrobat Reader, a freeware program, on your computer to view PDF files. See the link to Information about Acrobat Reader on the class web page. This has some suggestions as to how to use it plus a link to a site where you can download Acrobat Reader if you don't have it.

If you have problems, either in installing MacAnova or using it, please call me or send me email. You might also check the Frequently Asked Questions about MacAnova on the class web page.

Reading (week of September 12-16) Read J&W Chapter 3**Written Assignment** (due in class, Wednesday September 14):

Before attempting these, carefully study *An Introduction to MacAnova* and the handouts *Example of the Use of MacAnova* and *Making Plots Using MacAnova*. Among other things, the handout illustrates computing means and covariance matrices using matrix operation, using macro `covar()` and functions `describe()` and `tabs()`, and computing eigenvalues and eigenvectors using `eigen()`.

Note: In the examples below, in the most up-to-date version of MacAnova (except the Classic Macintosh version), you type the commands to the right of `Cmd>` in the lower panel of the MacAnova window, hit Enter or Return and what you typed appears in the upper part of the window preceded by `Cmd>` and followed by the output from the command.

1. Define the matrices

$$\mathbf{A} = \begin{bmatrix} 7 & 5 & 3 \\ 2 & 1 & 8 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 11 & -7 & 8 \\ 12 & 0 & 9 \end{bmatrix}$$

Statistics 5401/8401 Assignment Sheet 1

In MacAnova, you can create **A** and **B** by

```
Cmd> a <- matrix(vector(7,2, 5,1, 3,8),2) # 2 = no of rows
Cmd> b <- matrix(vector(11,12, -7,0, 8,9),2)# 2 = no of rows
```

Note that the data are entered column by column.

Use MacAnova to find the following matrices if possible. If not possible, state why. MacAnova expressions are in parentheses.

(a) A'	(a')	(f) A + B	(a + b)
(b) A - B	(a - b)	(g) A' B	(a %c% b or a' %*% b)
(c) AB	(a %*% b)	(h) A B'	(a %C% b or a %*% b')
(d) A'A	(a %c% a or a' %*% a)	(i) 17.3 A	(17.3 * a)
(e) A A'	(a %C% a or a %*% a')	(j) (1/19)A	(a/19 or (1/19)*a)

Note: You can use `help()` and `usage()` to get help on a command or other topic. Or **Help** on the **Help** menu will give you HTML based help.

2. Do Ex. 1.14 of J&W.

The data are in data set T01_06 (Table 1.6) in file JWdata5.txt. The data file has 98 rows of 6 columns. See **Retrieving Data from JWdata5.txt** below for information on how to retrieve data from this file.

In T01_06 the subject number is omitted. In its place is a column indicating the two groups. 1 indicates the non-multiple-sclerosis (non-MS) group and 2 indicates the multiple sclerosis (MS) group. Here's how you can extract non-MS group (excluding the indicator column 1) from the data:

```
Cmd> data <- read("", "T01_06") #see below
T01_06      98      6 format
) Data from Table 1.6 p. 42 in
) Applied Multivariate Statistical Analysis, 5th Edition
) by Richard A. Johnson and Dean W. Wichern, Prentice Hall, 2002
) These data were edited from file T1-6.DAT on disk from book
) by making making column 1 be group number (1 or 2) and removing
) last column which was group coded as 0 or 1.
)
) Multiple-Sclerosis data derived from evoked responses to visual
) stimuli recorded from the scalp of a subject. Two different
) visual stimuli S1 and S2 produced responses in the left (L) and
) right (R) eyes of subjects, some of whom had multiple sclerosis
) (MS).
) Col. 1: Group number, 1 = Non-MS group, 2 = MS group
) Col. 2: x1 = Age of subject
) Col. 3: x2 = S1L + S1R (Sum of left and right responses to
)          stimulus S1)
) Col. 4: x3 = abs(S1L - S1R)
) Col. 5: x4 = S2L + S2R
) Col. 6: x5 = abs(S2L - S2R)
Read from file "TP1:Stat5401:Data:JWData5.txt"

Cmd> groups <- data[,1]# Col. 1 identifies nonMS (1) or MS (2)

Cmd> unique(groups) # all the distinct values in groups
(1)                1                2
```

Statistics 5401/8401 Assignment Sheet 1

```

Cmd> data <- data[,-1]# eliminates col. 1
Cmd> list(data) # data now has 5 columns
data          REAL    98    5
Cmd> nonms <- data[groups == 1,] #selects nonMS cases
Cmd> ms <- data[groups == 2,]    #selects MS cases
Cmd> list(ms,nonms)
ms            REAL    29    5
nonms         REAL    69    5

```

The first command line read the data set using `read()` (equivalent to `matread()`). I next copied column 1 of data to a new variable `groups` and verified the values in `groups` are either 1 or 2. Then I deleted column 1 from data using a negative subscript. In the next two lines, I selected from data the rows (1st subscript) for which `groups` is 1 or 2, that is the data for the non-MS and the MS group, respectively. `list()` shows `ms` is a 29 by 5 matrix and `nonms` is 69 by 5 matrix.

Note the convention that *variables correspond to columns*, and *rows correspond to cases* or observations. This is assumed in all computations.

The remaining problems work with the data in J&W Table 11.5 or data set T11_05 in file `JWData5.txt`. These data consist of measurements on 4 numerical characteristics of 50 blossoms from each of three iris varieties, *I. Setosa* (cases or rows 1-50), *I. versicolor* (cases 51-100), and *I. virginica* (cases 101-150). These data were used by R. A. Fisher in a pioneering paper on discriminant analysis in taxonomy and are often referred to simply as the "Fisher Iris data."

Column 1 contains the variety number (1 – 4) and columns 2 through 5 contain, respectively, measurements on *sepal length*, *sepal width*, *petal length*, and *petal width*. Again, note that variables correspond to columns, and rows to cases or observations.

You can read the data in Johnson and Wichern Table 11.5 by

```
Cmd> irisdata <- read("", "T11_05")
```

You might verify that the data have the form described by typing

```
Cmd> irisdata[vector(1,2,49,50, 51,52,99,100,\
101,102,149,150), ]
```

This will print the first two cases and last two cases from each variety.

3. Let **X** be the the 50 by 4 matrix of *Iris setosa* data, with each variable in a *column*. Once you have read in T11_05, you can obtain **X** by

```
Cmd> variety <- irisdata[,1]; irisdata <- irisdata[,-1]
Cmd> x <- irisdata[variety == 1,] # selects setosa cases
```

Compute the the following using *MacAnova matrix operations only* (do not use `describe()`, `sum()`, `tabs()` or macro `covar()` except to check your answers):

- (a) The sample mean vector \bar{x}' (a 1 by 4 matrix or *row* vector; note the transpose symbol). You can check this by

```
Cmd> describe(x, mean:T)
```

- (b) The 4 by 4 matrix \mathbf{A} of mean corrected sums of squares and products

$$\mathbf{A} = \sum_{i=1}^{50} (x_i - \bar{x})(x_i - \bar{x})' = \left[\sum_{i=1}^{50} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_j) \right]_{1 \leq j, k \leq 4}.$$

- (c) The sample unbiased covariance matrix $\mathbf{S} = [s_{ij}]_{1 \leq j, k \leq 4} = (1/49)\mathbf{A}$

(Note $49 = n - 1 = 50 - 1$). You can check the correctness of the diagonal elements by `describe(x, var:T)` or the whole matrix by `tabs(x, covar:T)`.

- (d) You can compute the sample correlation matrix $\mathbf{R} = [r_{ij}]_{1 \leq j, k \leq 4}$, $r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}}$

as $\mathbf{R} = \mathbf{D} \mathbf{S} \mathbf{D}$, where

$$\mathbf{D} = \text{diag}[1/\sqrt{s_{ii}}]_{1 \leq i \leq 4} = \begin{bmatrix} 1/\sqrt{s_{11}} & 0 & 0 & 0 \\ 0 & 1/\sqrt{s_{22}} & 0 & 0 \\ 0 & 0 & 1/\sqrt{s_{33}} & 0 \\ 0 & 0 & 0 & 1/\sqrt{s_{44}} \end{bmatrix}.$$

This should give the same result as `cor(x)`.

4. Do the following further computations on the same data as in 3.

- (a) Using MacAnova operation `eigen()`, compute the eigenvalues (characteristic values) $\lambda_1, \lambda_2, \lambda_3$, and λ_4 and eigenvectors (characteristic vectors) $\mathbf{u}_1, \dots, \mathbf{u}_4$ of \mathbf{S} .
- (b) Let \mathbf{U} be the 4 by 4 matrix $[\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4]$ (that is, the matrix whose columns are the eigenvectors of \mathbf{S}), and let

$$\mathbf{D} = \text{diag}[\lambda_1, \lambda_2, \lambda_3, \lambda_4] = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{bmatrix}$$

Use operation `dmat()` to create \mathbf{D} from the output of `eigen()`. Show *numerically* that, except for possible rounding error, $\mathbf{U} \mathbf{D} \mathbf{U}' = \mathbf{S}$ and $\mathbf{U}' \mathbf{U} = \mathbf{U} \mathbf{U}' = \mathbf{I}_4$ (4 by 4 identity matrix with 1's on diagonal and 0's elsewhere).

5. Create a new 50 by 5 matrix \mathbf{Y} whose first four columns are the same as \mathbf{X} and whose last column is (petal length) + (petal width), that is, the sum of columns 3 and 4 of \mathbf{X} . One way to compute this is

```
Cmd> y <- hconcat(x, x[, 3] + x[, 4])
```

Alternatively, and more in the spirit of using matrix operations, $\mathbf{Y} = \mathbf{X} \mathbf{C}$, and can

therefore be computed by `y <- x %*% c`, where `c` contains matrix

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

The columns of \mathbf{C} define linear combinations of the columns of \mathbf{X} and the columns of \mathbf{Y} are linearly dependent since the last column is the sum of columns 3 and 4. You can create `c` by

```
Cmd> c <- matrix(vector(1,0,0,0, 0,1,0,0, 0,0,1,0, 0,0,0,1,\n0,0,1,1),4)
```

Note that elements are entered going down columns. Alternatively you might create `c` by

```
Cmd> c <- hconcat(dmat(4,1),vector(0,0,1,1))
```

since `dmat(4,1)` computes the 4 by 4 identity matrix. Look at the help for `hconcat()` and `dmat()`.

- Compute the sample (unbiased with denominator $n - 1$) covariance matrix \mathbf{S}_Y of \mathbf{Y} and its eigenvalues and eigenvectors. Note that the smallest eigenvalue is 0 (except possibly for rounding error) and the corresponding eigenvector consists of weights w_1, w_2, \dots, w_5 defining a linear combination of the columns of \mathbf{Y} which has zero variance. It should be a multiple of $\mathbf{c} = [0, 0, 1, 1, -1]'$ since $\mathbf{Y}\mathbf{c} = \mathbf{0}$. This illustrates that the eigenvectors corresponding to small eigenvalues of \mathbf{S}_Y help uncover linear dependences among data variables.
- Show *numerically* using MacAnova matrix `c` that the sample covariance matrix \mathbf{S}_Y of \mathbf{Y} can also be computed as $\mathbf{S}_Y = \mathbf{C}'\mathbf{S}_X\mathbf{C}$

Retrieving Data from JWData5.txt

You can read in the Table 1.6 data from file `JWdata5.txt` by

```
Cmd> data <- read("", "t01_06")
```

and then finding and selecting `JWData5.txt` in the dialog box. (Note: on Windows 95/98/NT/XP, the default options causes the "extension" `.txt` to be omitted in the dialog box so that the name of the file may appear to be `JWData5`.)

Alternatively, if you have set CHARACTER variable `DATAFILE` so that it has value the full name of the file, including the path, simply

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

may retrieve the data set. The simplest way to set `DATAFILE` is:

```
Cmd> DATAFILE <- getfilename()
```

`getfilename()` brings up a file navigation dialog box which you use to find and then select `JWData5.txt`. Since `getdata()` reads from the file whose name is in variable `DATAFILE`, after you do this, `getdata()` will always read from

Statistics 5401/8401 Assignment Sheet 1

JWData5.txt.

Alternatively, pre-defined macro `adddatapath()` adds a directory or folder to a list of places MacAnova will look for files. The simplest way to do this is to use `getfilename()` to find a folder:

```
Cmd> adddatapath(getfilename(pathonly:T))
```

In the file navigation dialog box, select any file in the folder or directory which you want to add to the list of places to search.

On the School of Statistics work station network, JWData5.txt is in directory `~kb/publicdata`. On a Statistics work station, you might use

```
Cmd> adddatapath("~kb/publicdata/")
```

to add this directory to the list of places to look for files.