

## BIO 499 Advanced Biometrics

### Week 8 Worksheet

**Note: This is completely voluntary. There is no extra credit for completing this worksheet. It will help you prepare for the final.**

#### Part One

We will work with a small data set consisting of observations for 24 smokers on 4 variables:

AGE: person's age

CIG\_DAY: number of cigarettes smoked per day

CO: a measure carbon monoxide level in the blood

MIN\_LAST: number of minutes since the person smoked their last cigarette

Below is a SAS program with the data to help you get started, and hopefully avoid data entry problems. You can get this "starter program" at the class web site.

```
dm 'output; clear; log; clear;';
options ls=95 ps=55 nodate pageno=1; *** FOR HARD COPIES - Portrait ***;
option formdlim = '_';
OPTIONS FORMCHAR="|----+|----+=|/\<>*";
data smoke;
input AGE CIG_DAY CO MIN_LAST;
datalines;
61 30 270 60
54 30 220 160
67 30 270 50
56 25 250 70
54 30 255 70
52 20 160 105
60 11 160 95
51 20 520 120
63 30 215 90
36 25 180 87
72 6 90 710
68 40 350 150
50 25 380 80
42 20 145 143
35 11 150 90
57 40 270 30
29 28 145 90
64 10 190 85
42 20 320 60
47 50 315 95
44 2 50 285
50 20 150 60
72 40 200 100
32 25 580 60
;
run;
```

Now, we'll have you add some SAS procedures and interpret their output, with the goal of learning some SAS and some statistics. If you wish, you can write your answers to the questions below directly on this document. You can then check your answers with the key available on the web site.

1. Utilize PROC PRINCOMP to do a Principal Components Analysis. Use the appropriate option of the PROC PRINCOMP statement to create an output SAS data set that contains all the original data as well as the principal component scores. Do not use any other options in this procedure. The first eigenvalue should be 1.83154557.

2. How much of the total variation in our four variables is explained by the first two components?

Provide an interpretation of the first component using the appropriate eigenvector.

3. Next put in PROC PRINT; Don't include any options or other statements with this procedure. Explain what the values are under PRIN1, PRIN2, PRIN3, and PRIN4. Don't just give the term that identifies what these numbers are, provide a short explanation of what they are. The first number under PRIN1 should be 0.53534.

4. Next put in PROC MEANS. Use the appropriate options in the PROC MEANS statement to print n, mean, and variance. (Hint: to find documentation on PROC MEANS, look in Base SAS (not in SAS/STAT), and then within Base SAS, look in Procedures.) The mean for PRIN1 should be 3.700743E-17.

Look at the means. What are those numbers *really*?

I mean, SAS (or Excel) calls it 3.700743E-17, but given the accuracy (number of decimal places) that SAS or Excel uses, what would we biologists call these numbers?

Now look at the variances. What are these numbers? Have you seen them before? What are these variances in the context of a Principal Components Analysis?

5. Next, put in PROC CORR. Use two options in the PROC CORR statement. One option should suppress displaying the probabilities associated with each correlation coefficient. The other option should suppress printing simple descriptive statistics for each variable. Use the VAR statement to only include variables PRIN1, PRIN2, PRIN3, and PRIN4. (Hint: documentation for PROC CORR is not in SAS/STAT. Go into Base SAS, and about 7 lines down you'll find documentation for PROCs CORR, FREQ, and UNIVARIATE.)

Your output from PROC CORR might remind you of the identity matrix. Explain why the correlations between the “variables” are so small.

6. Finally, put in another PROC PRINCOMP. This time, don't use any options in the PROC PRINCOMP statement. However, do include the VAR statement to limit the analysis to variables PRIN1, PRIN2, PRIN3, and PRIN4. The first eigenvalue should be 1.00000000.

Explain and interpret the eigenvalues, proportions, and eigenvectors here. Why are these numbers the way they are? (Hint: Remember what the values in PRIN1, PRIN2, PRIN3, and PRIN4 are.) Remember that we got these from the first PROC PRINCOMP. If you think about what these values are, it should help you explain the rather unusual values that come out of the second PROC PRINCOMP.

## Part Two

Let  $X$  be a square matrix of order 2, such that:  $X = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}$

1. Write the characteristic equation for matrix  $X$ .
2. Now, solve the characteristic equation to determine the eigenvalues of  $X$ . You must show your work, and show what you're doing at each step. You are welcome to use PROC IML (or any other computer tool at your disposal) to see what are the correct eigenvalues, but to get credit you must show you calculated them from the characteristic equation.
4. What is the trace of  $X$ , i.e.  $\text{tr } X$ ?
5. What is the sum of your eigenvalues?
6. How does the sum of the eigenvalues relate to  $\text{tr } X$ ? Are you surprised?
7. Write (but do not solve!) the equation involving  $X$  that is used to determine the eigenvectors of  $X$ .

### Part Three

Oh, goody! A PCA game! Data are for six patients admitted to a hospital. The variables are:

Dur\_stay: duration of their stay in the hospital (in days)

Age: the age of the patient

Temp: the first body temperature measurement following their admission to the hospital

WBC: first white blood cell counts following admission (if you're a medical doctor, you realize the numbers are actually x1000, i.e. 8 is really 8000. Use the numbers as they are presented here, don't multiply them by 1000.)

Data				
Patient	Dur_stay	Age	Temp	WBC
A	5	30	99	8
B	10	73	98	5
C	6	40	99	12
D	11	47	98.2	4
E	5	25	98.5	11
F	14	82	96.8	6

Below is a SAS program with the data to help you get started, and hopefully avoid data entry problems. You can get this "starter program" at the class web site.

```
dm 'output; clear; log; clear;';
*options ls=76 ps=55 pageno=1; *** FOR THE SCREEN ***;
*options ls=123 ps=41 nodate pageno=1; *** FOR HARD COPIES - Landscape ***;
options ls=95 ps=55 nodate pageno=1; *** FOR HARD COPIES - Portrait ***;
option formdlm = '_';
OPTIONS FORMCHAR="|----|+|----+=|-\<>*";
Title 'PCA Game - Hospital Data';
Data Hospital;
input Patient $ Dur_stay Age Temp WBC;
datalines;
A 5 30 99 8
B 10 73 98 5
C 6 40 99 12
D 11 47 98.2 4
E 5 25 98.5 11
F 14 82 96.8 6
;
run;
```

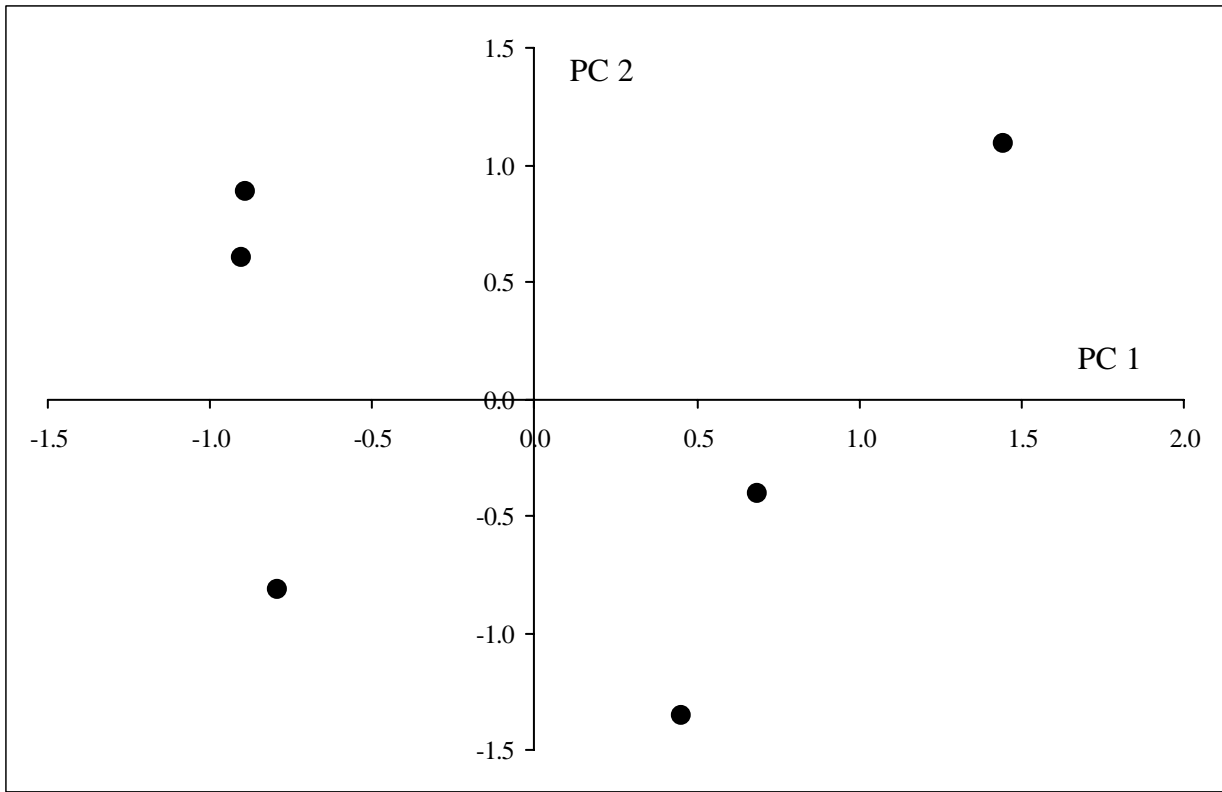
#### Correlation Matrix

	Dur_stay	Age	Temp	WBC
Dur_stay	1.00			
Age	0.89	1.00		
Temp	-0.91	-0.85	1.00	
WBC	-0.76	-0.63	0.57	1.00

#### Eigenvalues

		Factor Pattern	
		Factor1	Factor2
1	3.31623566		
2	0.48449706	Dur_stay	0.98033 0.02890
3	0.14900257	Age	0.93200 0.19059
4	0.05026471	Temp	-0.92191 -0.29578
		WBC	-0.79790 0.59988

On the next page is the graph of the six patients plotted by their factor scores on PC 1 and PC 2.



Your task is to identify each patient. Write the letter for each patient by the correct point.

Now, do the PCA on SAS (using PROC FACTOR or PROC PRINCOMP; your choice). Use the PCA to determine the correct identity for each patient, and see how you did! If you've guessed wrong, try to explain why your guess was wrong. If you wish, write your explanations here.