S&DS 363/563 and ENV 758
Multivariate Statistics Homework #2
Principle Components Analysis
Due : Thursday, 2.8.24 11:59pm on CANVAS

Be sure to check out sample programs in R and SAS on the CANVAS home page

Answers should be complete and concise. You should turn in typed solutions. If you are working in a group, you may turn in one problem set per group (list all group members). You may use any statistics program for calculations that you wish. If you use SAS or R, please include your code (either as it's relevant or at the end). You can also use R Markdown and submit both a knitted PDF or Word Doc as well as you .RMD code. If using SAS, you can submit a separate .SAS file if you like.

SAMPLE DATA SET

The example below is JUST FOR YOUR PRACTICE. NOTHING TO TURN IN HERE!

The data set `AirPollution.xls` is an excel file that contains weather/pollution measurements on 42 consecutive days at one site in Los Angeles. Each day, measurements were taken at precisely 12 noon. There are seven variables:

Wind
Solar Radiation
Carbon Monoxide
Nitrogen Oxide
Nitrogen Dioxide
Ozone
Hydrogen Chloride

Your goal is to see if these measurements can be summarized in fewer than seven dimensions.

1). Compute the correlation matrix between all variables (SAS and SPSS will provide this for you as part of the PCA procedure – in SPSS, click on DESCRIPTIVES, in R use the cor() function or any of the cool correlation plots discussed in class.). Comment on relationships you do/do not observe.

Correlation Matrix

		Wind	Radiation	CO	NO	NO2	O3	HC
Correlation	Wind	1.000	-.101	-.194	-.270	-.110	-.254	.156
	Radiation	-.101	1.000	.183	-.074	.116	.319	.052
	CO	-.194	.183	1.000	.502	.557	.411	.166
	NO	-.270	-.074	.502	1.000	.297	-.134	.235
	NO2	-.110	.116	.557	.297	1.000	.167	.448
	O3	-.254	.319	.411	-.134	.167	1.000	.154
	HC	.156	.052	.166	.235	.448	.154	1.000

There are some relationships – mostly between CO and other Oxygen-containing compounds.

2). Perform Principle components analysis using the Correlation matrix (standardized variables). Think about how many principle components to retain. To make this decision look at

- Total variance explained by a given number of principle components
- The ‘eigenvalue > 1’ criteria
- The ‘scree plot elbow’ method
- Parallel Analysis : for the air pollution data, the first five threshold values for the Allen and Longman methods are provided below (based on n=42 observations, p=7 variables) :

```
eigenval    LONGMAN     ALLEN
1           1.77411     1.78971
2           1.44221     1.52097
3           1.22756     1.32395
4           1.02647     1.16865
5           0.89682     1.03550
```

As you make this decision, keep in mind that the number of observations is somewhat small relative to the number of variables.

Here are SPSS results :

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.337	33.383	33.383	2.337	33.383	33.383
2	1.386	19.800	53.183	1.386	19.800	53.183
3	1.204	17.201	70.384	1.204	17.201	70.384
4	.727	10.387	80.771	.727	10.387	80.771
5	.653	9.335	90.106	.653	9.335	90.106
6	.537	7.667	97.773	.537	7.667	97.773
7	.156	2.227	100.000	.156	2.227	100.000

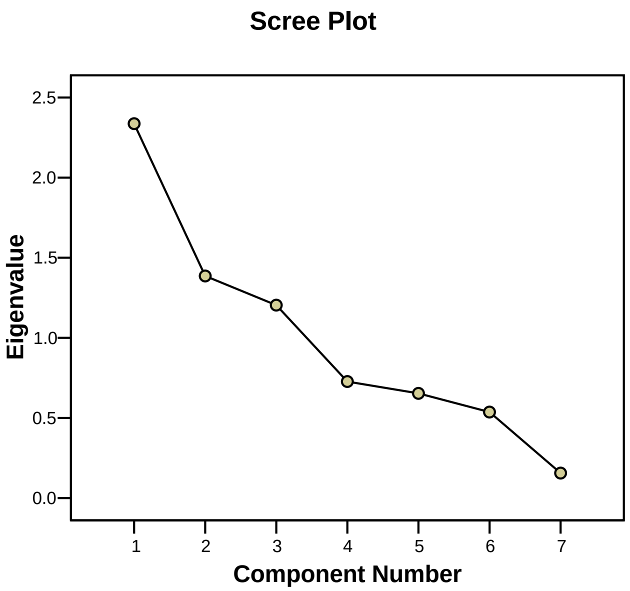
Extraction Method: Principal Component Analysis.

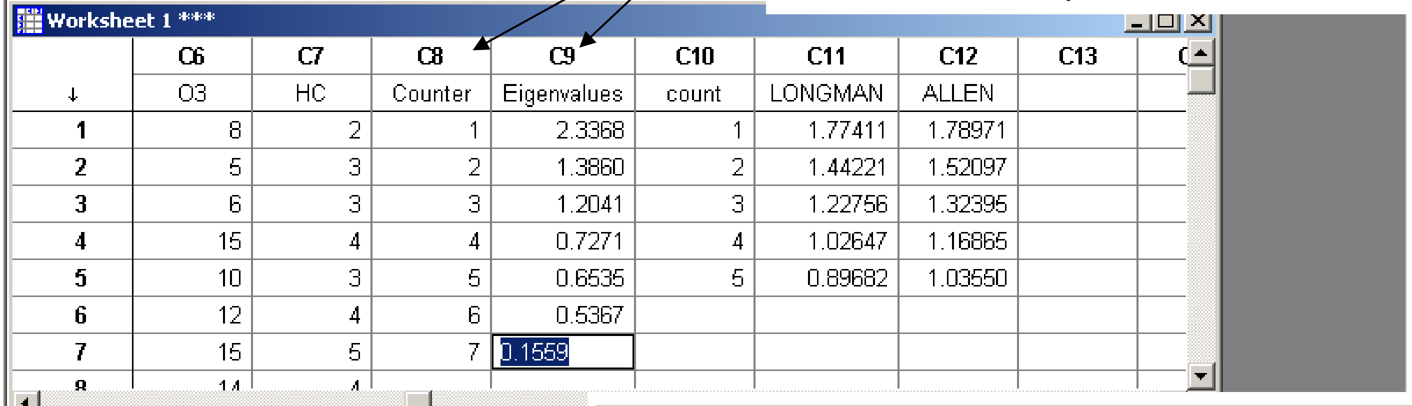
An 80% threshold would argue for 4 components. The eigenvalue greater than 1 rule would argue for 3 components.

Scree plot is sort of double-jointed – elbow at two and four, which would argue for retaining one or three components.

For Parallel analysis, you can use SAS, R, or the SPSS Macro online. In MINITAB, do the following using the data provided in the table above.

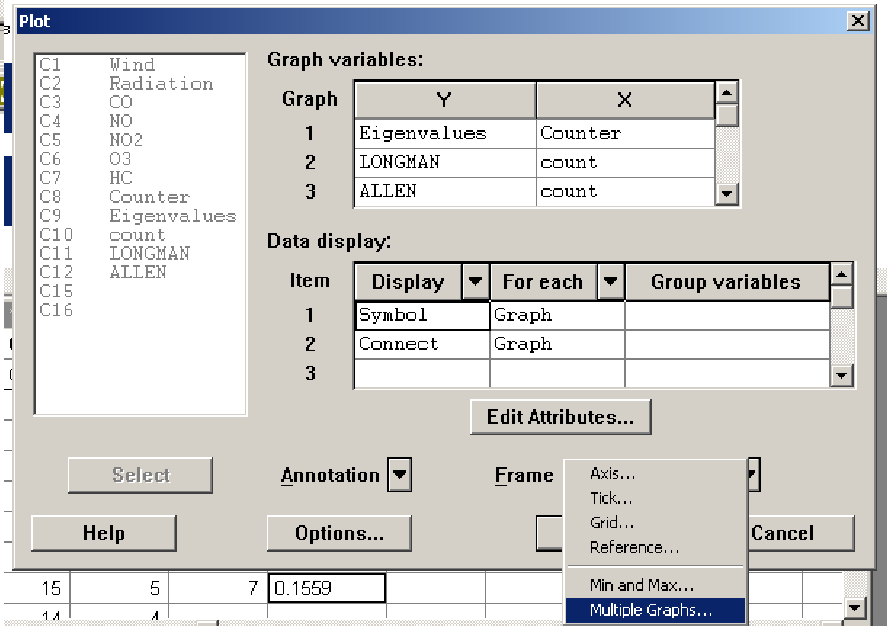
1) *Copy the data above into MINITAB.*
2) *Make two more variables – one which has the eigenvalues calculated by MINITAB (copy from output screen), one which is a counter for the eigenvalue number (here from 1 to 7)   : see below*





↓	C6 O3	C7 HC	C8 Counter	C9 Eigenvalues	C10 count	C11 LONGMAN	C12 ALLEN	C13
1	8	2	1	2.3368	1	1.77411	1.78971	
2	5	3	2	1.3860	2	1.44221	1.52097	
3	6	3	3	1.2041	3	1.22756	1.32395	
4	15	4	4	0.7271	4	1.02647	1.16865	
5	10	3	5	0.6535	5	0.89682	1.03550	
6	12	4	6	0.5367				
7	15	5	7	0.1559				

Current Worksheet: Worksheet 1



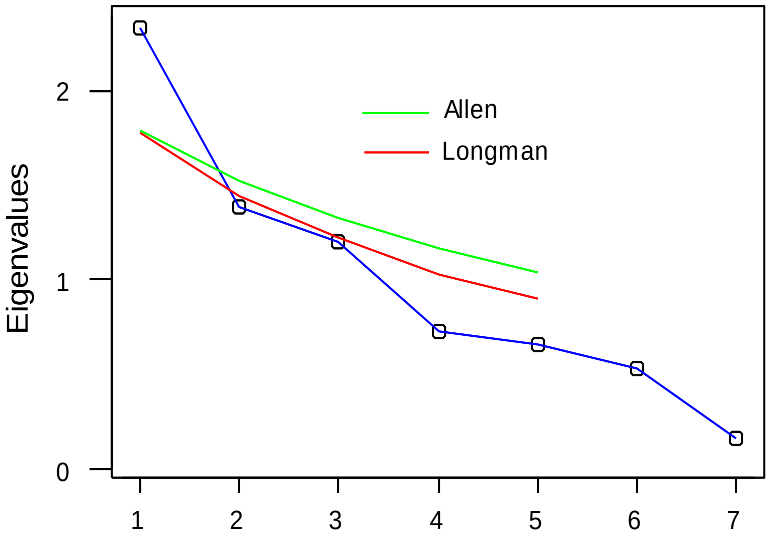
Graph variables:

Graph	Y	X
1	Eigenvalues	Counter
2	LONGMAN	count
3	ALLEN	count

Data display:

Item	Display	For each	Group variables
1	Symbol	Graph	
2	Connect	Graph	
3			

3) *Under* Graph → Plot, *input three Y,X combinations : Eigenvalues vs. Counter, Longman vs. count, Allen vs. Count. Under Frame, choose multiple graphs and indicate that the plots should be overlayed. Indicate that Symbols and Connect (i.e. lines) should be displayed for each plot. Use Edit Attributes to change colors, plot characters, etc.*



Both parallel methods suggest retaining one principle component. Be aware that since the number of observations is small (only 44), the parallel procedures will more easily reject components with borderline eigenvalues. Also keep in mind that the parallel procedure assumes the variables have a normal distribution, a bit questionable here. I decide to keep three components since with three components, I can explain 70% of the variability in the data (only 33% of variability with one component – however, keeping one component is also a reasonable decision).

3). For principle components you decide to retain, examine the loadings (principle components) and think about an interpretation for each component.

```
Variable           PC1      PC2      PC3
Wind             0.237   -0.278    0.643
Radiatio        -0.206    0.527    0.224
CO              -0.551    0.007   -0.114
NO              -0.378   -0.435   -0.407
NO2             -0.498   -0.200    0.197
O3              -0.325    0.567    0.160
HC              -0.319   -0.308    0.541
```

Component one is mostly CO, NO2. Component 2 is Radiation, NO, and Ozone. Component 3 is Wind, NO, HC. The division is not exact. However, with three measures, I can explain 70% of the variability.

4). Write a paragraph summarizing your findings, and your opinions about the effectiveness of using principle components on this data.

Not being a weather expert, I can't say much about the interpretation of the factors beyond what was stated above. Given the relatively small number of observations, principle components was not entirely successful. Interpretations of the factors is somewhat difficult.

HOMEWORK ASSIGNMENT

PLEASE turn in the following answers for YOUR DATASET! If PCA is not appropriate for your data, use ONE of the datasets online (either DrugAttitudes.xls or Pizza.csv described on the following pages).

List your name and a one sentence reminder of which dataset your are using.

1). First, discuss whether your data seems to have a multivariate normal distribution. Make univariate plots (boxplots, normal quantile plots) as appropriate – just make it clear you've gotten familiar with the data. Then make transformations as appropriate. You do NOT need to turn all this in, but describe what you did. **THEN** make a chi-square quantile plot of the data. Turn in your chi-square quantile plot and comment on what you see. **NOTE that multivariate normality is NOT a requirement for PCA to work!**

2). Compute the correlation matrix between all variables (SAS and SPSS will provide this for you as part of the PCA procedure – in SPSS, click on DESCRIPTIVES. In R use the cor() function or one of the other cool correlation plots.). Comment on relationships you do/do not observe. Do you think PCA will work well?

3). Perform Principle components analysis using the Correlation matrix (standardized variables). Think about how many principle components to retain. To make this decision look at:

- Total variance explained by a given number of principle components
- The 'eigenvalue > 1' criteria
- The 'scree plot elbow' method (turn in the scree plot)
- Parallel Analysis: think about whether this is appropriate based on what you discover in question 1.

4). For principle components you decide to retain, examine the loadings (principle components) and think about an interpretation for each retained component if possible.

5) Make a score plot of the scores for at least one pair of component scores (one and two, one and three, two and three, etc). Discuss any trends/groupings you observe (probably, this will be 'none'). **As a bonus, try to make a 95% Confidence Ellipse for two of your components.** You might want to also try making a bi-plot if you're using R.

6). Write a paragraph summarizing your findings, and your opinions about the effectiveness of using principle components on this data. Include evidence based on scatterplots of linearity in higher dimensional space, note any multivariate outliers in your score plot, comment on sample size relative to number of variables, etc.

LOANER DATASETS
(if PCA is not appropriate for your data)

Drug Attitudes

The data set `DrugAttitudes.xls` is an excel file that contains attitudes of 38 people measured on 20 variables relating to drugs. Each question was measured on a 5 point scale where 1=Strongly Agree and 5 = Strongly Disagree. The variables were

legal	All drugs should be made legal and freely available.
dangerous	As a general rule of thumb, most drugs are dangerous and should be used only with medical authorization.
regret	Drugs can cause people to say or do things they might later regret.
unnatural	Drugs are basically an "unnatural" way to enjoy life.
notuse	Even if my best friend gave me some hash, I probably wouldn't use it.
psycho	Experimenting with drugs is dangerous if a person has any psychological problems.
trip	I see nothing wrong with taking an LSD trip.
stoned	I admire people who like to get stoned.
calm	I wish I could get hold of some pills to calm me down whenever I get "up tight".
high	I would welcome the opportunity to get high on drugs.
noaspirin	I'd have to be pretty sick before I'd take any drug including an aspirin.
relationship	If people use drugs together, their relationships will be improved.
drugscene	In spite of what the establishment says, the drug scene is really "where it's at".
caregivers	People who regularly take drugs should not be given positions of responsibility for young children.
experience	People who make drug legislation should really have personal experience with drugs.
fun	People who use drugs are more fun to be with than those who don't use drugs.
stupid	Pep pills are a stupid way of keeping alert when there's important work to be done.
lessalcohol	Smoking marijuana is less harmful than drinking alcohol.
sideeffects	Students should be told about the harmful side effects of certain drugs.
dope	Taking any kind of dope is a pretty dumb idea.

Your goal is to see if these measurements can be summarized in fewer than 20 dimensions. **NOTE that one variable may get imported as a text variable – this might cause you problems.**

Nutrients in Pizza

The data set `Pizza.csv` contains nutrient information for a number of brands of pizza. Multiple samples were taken from each brand.   The variables are:

brand	Pizza brand (class label)
id	Sample analyzed
mois	Amount of water per 100 grams in the sample
prot	Amount of protein per 100 grams in the sample
fat	Amount of fat per 100 grams in the sample
ash	Amount of ash per 100 grams in the sample
sodium	Amount of sodium per 100 grams in the sample
carb	Amount of carbohydrates per 100 grams in the sample
cal	Amount of calories per 100 grams in the sample

When you make a scoreplot, you may want to try to save the scores, and make a plot where you use different colors/symbols for each brand of pizza.   Another thing you may want to try is to analyze the data by first excluding brand A.