S&DS 3630/5360/ENV 758
Multivariate Statistics
Homework #2: Principle Components Analysis

Due: 2/3/26 11:59pm on CANVAS

Be sure to check out sample programs in R and SAS on the CANVAS home page. Examples in SPSS are available in the notes.

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) ONLY if you are using your own dataset. If you are using the loaner data, each person must turn in their own problem set. You may use any statistics program for calculations that you wish. If you use SAS or R, please include your code (either when relevant or at the end). You can also use R Markdown and as a PDF or a knitted Word document.

PLEASE turn in the following answers for YOUR DATASET! If PCA is not appropriate for your data, use ONE of the datasets described at the end (all available up on CANVAS).

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

1) First, provide some plots showing relationships between your variables (i.e. scatterplots, etc). Discuss what you see, thinking in particular about high-dimensional linearity. Next, examine the multivariate normality of your data – make plots as appropriate including a chi-square quantile plot. If you decide to make transformations of variables, make at least some post-transformation plots (including a new chi-square quantile plot) and again discuss linearity and multivariate normality. **NOTE that multivariate normality is NOT a requirement for PCA to work!**

2) Examine the correlations among all of your variables. Include your results in table/graph form as you deem appropriate. Comment on how well you think PCA will work on your data. In addition, provide a discussion of sample size relative to the number of variables in your dataset.

3) Perform Principal components analysis using the Correlation matrix (standardized variables). Think about how many principal 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 principal components you decide to retain, examine the loadings (principal components) and think about an interpretation for each retained component.

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'). In addition, make a 95% CI ellipse for two of the retained components. Discuss whether it makes sense to use this as an outlier detection method and describe what you observe. If possible, include a bi-plot as well and discuss what you observe.

6) Write a paragraph summarizing your findings and your opinions about the effectiveness of using PCA on your data. Include evidence based on scatterplots of linearity in higher dimensional space, note any multivariate outliers in your score plot, interpretation of components, etc.

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

Drug Attitudes

The data set `DrugAttitudes.csv` 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.

Air Pollution

The data set `AirPollution.csv` 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