STA280 Unsupervised Learning
Homework on Principal Components Analysis

Directions: *Be sure to show your work and explain your answer for each question, even if the question seems to require only a Yes or No answer. Your homework solutions are to be entirely your own effort. You may not communicate with anyone about the homework, except for the TA and/or the instructor. You may use the Canvas postings, in-class discussion, any of the recommended textbooks, and computer software, if necessary, but no other resources. In writing up your solutions, it is recommended to support your answers with "cut-and-pasted" output, provided your answers are clearly* labeled *and* circled *or* highlighted*. The grader will not search through unlabeled computer output to try to find your answers.*

Context for these homework questions:

A management consulting firm studied the requirements of each position filled by 67 employees of a firm and assigned numerical ratings (real data) to three job dimensions that the consulting firm thought important. These dimensions are:

- the knowhow demanded by the job
- the level of problem-solving skill required by the job
- the amount of accountability inherent in the job.

The ratings are evaluations of the jobs, not of the particular employees who actually fill the positions. The data are in the text file "**job ratings.txt**" and include a job code (not further identified) and the annual salary in dollars of the person actually occupying each position, in addition to the ratings.

High-level hints: You will need to write a SAS program and/or run JMP analyses to answer the homework questions. In order to minimize your work, you may want to read over all of the questions first to plan your program and analyses rather than taking one question after another piece-meal. Some of the questions suggest that you may want to do some calculations in Excel. This is to minimize your effort. If you do Excel computations, submit your Excel file showing your work and leaving your formulas in the cells. This will help resolve potential grading issues that would arise if you convert to Word or PDF – both of which would lose the formulas.

1. Create a SAS dataset called **WORK.RATINGS** that contains the data in the **job ratings.txt** file. Assign the SAS names **JOB, KNOWHOW, PROBLEM_SOLVING, ACCOUNTABILITY, SALARY**, respectively, to the five variables as they appear from left to right in the file. Extract the principal components of the three dimensions that were rated by the management consulting firm. Use the default (standardized) version of the extraction. *Your answer for question 1 is your SAS code only.*

2. This question verifies the basic property of principal components transformations.
   a) Write the equations of the principal components of the PCA in question 1.
   b) Verify that the principal component transformation in question 1 is an orthonormal rotation of the (standardized) original three dimensions by showing that the rotation matrix satisfies the definition of an orthonormal transformation.

[Hint: You may find it helpful to perform the computations in Excel. You may wish to submit Excel computations as your solution.]

3. This question partially verifies the geometry-preserving property of principal components transformations.

a) Rotate the first two jobs in the text file by calculating their principal component scores.
b) The rotated scores for the two jobs in part (a) are each a vector of three scores. Verify that the lengths of these two vectors are the same as the lengths of the original (but standardized) ratings vectors of the two jobs.
c) Verify that the angle between these two rotated vectors is the same as the angle between the original unrotated vectors.

[Hint: You may find it helpful to perform the computations in Excel. You may wish to submit Excel computations as your solution.]

4. Obtain the principal components scores for all 67 jobs. Calculate the variances of the three sets of scores and verify that the variances are equal to the eigenvalues of the PC transformation. *[Hint: If you want to calculate variances in Excel, you can output the PC scores into a SAS dataset in SAS OnDemand and then convert it into a downloadable Excel file using the code in this footnote.*[1] *You may also get the scores in **JMP** and then copy into Excel.]*

5. Find the regression equation that results from regressing **PRIN1** on the three ratings knowhow, problem_solving, and accountability after the ratings have been standardized and without an intercept.[2] Are you surprised by the equation? *[Hint: You can standardize the variables manually – say in Excel – and then enter them into a SAS dataset for regression. Alternatively, you can standardize in SAS by using (e.g.)* ***PROC STDIZE DATA=TWS.APTS OUT=TWS.APTS_STD; VAR X1 X2;*** *The variables X1, X2, … will be* replaced *by their standardized versions in the newly-created SAS dataset TWS.APTS_STD .]*

6. Find the regression equation that results from regressing (standardized) **KNOWHOW** on the three principal components without an intercept. Are you surprised by the equation? *[Hint: see preceding hint.]*

7. Write the **loadings matrix**, structured with components as columns and variables as rows. Using the loadings matrix, try to interpret meanings for the three principal components. *[Hint: SAS does not provide the loadings matrix directly in **PROC PRINCOMP**. But it can be obtained as a correlation table from **PROC CORR**, or from the eigenvectors and their standard deviations. Check the various save icons for SAS OnDemand results. Alternatively, the loadings matrix can be obtained directly from JMP.]*

8. How many principal components would you retain …

a) Using the Kaiser rule?
b) Using the Joliffe rule?
c) Using the 80% rule?

---

[1] For example, PROC EXPORT DATA=TWS.apts OUTFILE="/home/tomsager/my_content/aptsexcel.xlsx" DBMS=XLSX REPLACE; RUN; is SAS code that converts a SAS dataset **TWS.apts** into an Excel file **aptsexcel.xlsx** that is placed in my "MyFolders/my_content" subfolder in SAS OnDemand. You cannot EXPORT / IMPORT directly between SAS OnDemand and your computer; you must go through the file structure in SAS OnDemand as an intermediary.

[2] To run a no-intercept regression, use the **NOINT** option (e.g.,) **PROC REG DATA=TWS.APTS; MODEL Y = X1 X2 / NOINT; RUN;**

9. Find the regression equation that results from regressing **salary** on the three principal components with intercept. How much explanatory power do the three PCs collectively have in explaining **salary**?

10. In terms of explaining **salary**…
    a) Which component is most useful? Second most useful? Least useful?
    b) Is the usefulness of the PCs for explaining salary in the order PC1 > PC2 > PC3?
    c) How much explanatory power is lost if one uses only PRIN1 to explain **salary**?