Unsupervised Learning HW2

Submitted By: Parthiv Borgohain (pb25347)

1. Create an R or JMP dataset called RATINGS that contains the data in the source data file (job ratings.txt or job ratings.xlsx). Assign the names JOB, KNOWHOW, PROBLEM_SOLVING, ACCOUNTABILITY, SALARY, respectively, to the five variables as they appear from left to right in the file. (These are the names that appear in the first line of each source file and should be the names automatically assigned to the variables under default file reading.) 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 R script and/or your JMP script code only. [Hint: If you use JMP, you can save the JMP script that your pointing-and-clicking generate by clicking the File menu on the JMP Home Page and then selecting Save Session Script … toward the bottom. Your JMP script will be saved as a text (Notepad-readable) file with the extension jsl.]

   Answer: Below is a program written in R to create a dataset RATINGS from job_ratings.txt:

```
# First, read in the data and set it to a dataset RATINGS

library(readxl)

RATINGS <- read_excel("Job Ratings.xlsx")

View(RATINGS)


# Extracting Principal Components on 3 selected variables by Consulting firm (scale=TRUE).

PCA.RATINGS <- prcomp(RATINGS[2:4], scale=TRUE)


# Names of variables in the PCA object

names(PCA.RATINGS)


#Interpretation of PCs:

PCA.RATINGS$rotation


# Relative importance and summary of the PCs

summary(PCA.RATINGS)
```

Here is the output obtained after running PCA:

```
> #Interpretation of PCs:
> PCA.RATINGS$rotation
                       PC1        PC2        PC3
knowhow         -0.5762506  0.6181215 -0.5346598
problem_solving -0.5843425  0.1457584  0.7983097
accountability  -0.5713835 -0.7724509 -0.2772013
```

2. This question verifies some 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) attributes (the original three dimensions) by showing that the rotation matrix satisfies the definition of an orthonormal transformation. [Hint: You may find it convenient to perform the computations in Excel. You may wish to submit Excel computations as your solution.]

Answer: Here are the Principal Components as obtained from R:

```
> #Interpretation of PCs:
> PCA.RATINGS$rotation
                       PC1        PC2        PC3
knowhow         -0.5762506  0.6181215 -0.5346598
problem_solving -0.5843425  0.1457584  0.7983097
accountability  -0.5713835 -0.7724509 -0.2772013
```

Hence, the equations are-

a)

PC1 = (-0.5762506*knowhow)+(-0.5843425*problem_solving)+(-0.5713835*accountability)

PC2 = (0.6181215*knowhow)+(0.1457584*problem_solving)+(-0.7724509*accountability)

PC3 = (-0.5346598*knowhow)+(0.7983097*problem_solving)+(-0.2772013*accountability)

b) Now we will paste the PCs onto Excel and compute the vector lengths and dot products. These are the results as obtained on excel-

Principal Components (from R)			
	PC1	PC2	PC3
knowhow	-0.5762506	0.6181215	-0.5346598
problem_solving	-0.5843425	0.1457584	0.7983097
accountability	-0.5713835	-0.7724509	-0.2772013

	Vector Lengths		
	PC1^2	PC2^2	PC3^2
	0.33206475	0.38207419	0.2858611
	0.34145616	0.02124551	0.63729838
	0.3264791	0.59668039	0.07684056
Sum	1.0000	1.0000	1.0000

Dot Product of Vectors	
Dot Product (PC1,PC2)	0.0000
Dot Product (PC1,PC3)	0.0000
Dot Product (PC2,PC3)	0.0000

So, the Principal Components are vectors of length 1. Also, their inner products with each other are 0. So clearly, the Principal Component transformation is an orthonormal transformation

Clearly, the **vector lengths of the PCs are 1**. Also, their **dot products with each other are 0 which** means that they are orthogonal to one another. Hence, **the transformation is an orthonormal transformation.**

3. This question partially verifies the geometry-preserving property of principal components transformations. a) Take the first two jobs in the text file and transform them from attribute space into component space 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 convenient to perform the computations in Excel. You may wish to submit Excel computations as your solution.]

Answer:

a) We first took the entire original dataset and standardized the columns knowhow,problem_solving,accountability. Then, we took the first two rows and transformed them into the Principal Components Space by using the equations described in Q2 a) on Excel. The results obtained on excel are shown below-

First two Jobs (From Original Standardized Dataset)			
job	knowhow	problem_solvin	accountability
0	4.35032492	5.283939792	6.116424944
2	2.25124045	2.122082877	2.124774933

First Two jobs in Principal Components Space			
job	PC1	PC2	PC3
0	-9.089332225	-1.265429978	0.196795593
2	-3.751363211	0.059567219	-0.098558798

Principal Components (from R)			
	PC1	PC2	PC3
knowhow	-0.5762506	0.6181215	-0.5346598
problem_solving	-0.5843425	0.1457584	0.7983097
accountability	-0.5713835	-0.7724509	-0.2772013

b) Then we computed the vector lengths for these 2 rows in both the original (standardized) feature space and in the Principal Components Space. These were the results obtained in Excel-

First two Jobs (From Original Standardized Dataset)				
job	knowhow	problem_solving	accountability	Length of Vectors
0	4.35032492	5.283939792	6.116424944	9.1791
2	2.25124045	2.122082877	2.124774933	3.7531

First Two jobs in Principal Components Space				
job	PC1	PC2	PC3	Length of Vectors
0	-9.089332225	-1.265429978	0.196795593	9.1791
2	-3.751363211	0.059567219	-0.098558798	3.7531

The first job has a **vector length of 9.1791** in **both the feature spaces** and the second job has a **vector length of 3.7531** in **both the spaces.** Clearly, the **length of the two vectors are equal.**

c) I computed the dot products between the two jobs in both the feature spaces (original and transformed Principal Components Space). The **Dot product in both the spaces was 34.0026**. Then using the I divided the dot product by the product of the vector magnitudes to find the cosine of the angle between the two vectors. The **cosines in both feature spaces was 0.9870.** Then we took the arccosine in order the compute the angle. The working done in Excel is shown below-

First two Jobs (From Original Standardized Dataset)				
job	knowhow	problem_solving	accountability	Length of Vectors
0	4.35032492	5.283939792	6.116424944	9.1791
2	2.25124045	2.122082877	2.124774933	3.7531

First Two jobs in Principal Components Space				
job	PC1	PC2	PC3	Length of Vectors
0	-9.089332225	-1.265429978	0.196795593	9.1791
2	-3.751363211	0.059567219	-0.098558798	3.7531

Principal Components (from R)			
	PC1	PC2	PC3
knowhow	-0.5762506	0.6181215	-0.5346598
problem_solving	-0.5843425	0.1457584	0.7983097
accountability	-0.5713835	-0.7724509	-0.2772013

Inner Dot Products between the two vectors	
In Original Space	34.0026
In PC Space	34.0026

Dot Product divided by Product of Vector Lengths	
In Original Space	0.9870
In PC Space	0.9870

Arc Cosine to get angle in radians	
In Original Space	0.1614
In PC Space	0.1614

So, the **angle between the two vectors is the same** in both the feature spaces. The angle between them in **both the cases is 0.1614 radians.**

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: Once you have the PCs, you can calculate variances within R or JMP. If you prefer to calculate variances in Excel, you can get the scores in JMP and then copy into Excel by clicking on the red down chevron just to the left at the top of the Principal Components: on Correlations window, then select Save Columns, then Save Principal Components, and the number of components that you specify will be added to your Ratings data window. You can then use the File menu to Save As an xlsx workbook.]

Answer:
The eigenvalues and eigenvectors of the PC transformation were obtained from R as shown below:

```
> eigen(m)
eigen() decomposition
$values
[1] 2.908081137 0.083697370 0.008221492

$vectors
           [,1]       [,2]       [,3]
[1,] -0.5762506 -0.6181215  0.5346598
[2,] -0.5843425 -0.1457584 -0.7983097
[3,] -0.5713835  0.7724509  0.2772013
```

For computing the variances in the 3 Principal Components, I used Excel. Firstly, I transformed every observation into the transformed Principal Components Space and then computed the variances of the Three Principal Components. The detailed working is available in the attached Excel File. Here are the final results-

Variances		
PC1	PC2	PC3
2.908081	**0.083697**	**0.008221**

Eigenvalues (Copied from R)	
1	**2.908081137**
2	**0.08369737**
3	**0.008221492**

Clearly, the variances of the PCs are equal to the eigenvalues of the PC transformation as seen in the above screenshot from Excel.

5. Find the regression equation that results from regressing PRIN1 on the three ratings knowhow, problem_solving, and accountability, without an intercept, 1 after the ratings have been standardized to mean 0 and variance 1. Are you surprised by the equation? [Hint: It may be convenient to standardize the variables manually – say in Excel – and then read them into R or JMP for regression.]

   Answer:
   I stored all the original standardized variables along with the transformed PC variables and stored them as one dataset in Excel. Then I imported this dataset into R. On this dataset, I ran a zero-intercept linear regression with PC1 as the dependent variable and knowhow,problem_solving and accountability as the dependent variables. These are the coefficients obtained after running the regression in R:

Linear Regression Model details copied from R

lmPC1	list [12] (S3: lm)	List of length 12
coefficients	double [4]	1.19e-09 -5.76e-01 -5.84e-01 -5.71e-01
(Intercept)	double [1]	1.194028e-09
knowhow	double [1]	-0.5762506
problem_solving	double [1]	-0.5843425
accountability	double [1]	-0.5713835
residuals	double [67]	-6.55e-10 6.76e-10 -3.15e-09 3.57e-09 4.12e-09 4.12e-09
effects	double [67]	-9.77e-09 1.36e+01 -2.30e+00 -1.14e+00 4.30e-09 4.30e
rank	integer [1]	4
fitted.values	double [67]	-9.09 -3.75 -3.21 -3.15 -2.98 -2.98 ...

   As seen in the screenshot above, the coefficients are:
   Knowhow: -0.5763, problem_solving: -0.5843, accountability: -0.5714

   So, the equation obtained is –
   PC1 = (-0.5763*knowhow)+(-0.5843*problem_solving)+(-0.5714*accountability)

   Clearly, the three **regression coefficients** are **exactly the same** as the **coefficients for PC1** as done in Q2. This is not surprising and is expected behavior as the coefficients are the same as the eigenvector values.

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.]

   Answer: Regression Output from R-

Linear Regression Model details copied from R

Name	Type	Value
lm_knowhow	list [12] (S3: lm)	List of length 12
coefficients	double [4]	6.27e-10 -5.76e-01 6.18e-01 -5.35e-01
(Intercept)	double [1]	6.27119e-10
PC1	double [1]	-0.5762506
PC2	double [1]	0.6181214
PC3	double [1]	-0.5346598
residuals	double [67]	-4.78e-10 3.12e-10 -1.60e-09 1.80e-09 2.58e-09 2.58e-09
effects	double [67]	1.07e-14 -7.98e+00 1.45e+00 3.94e-01 2.60e-09 2.60e-0..
rank	integer [1]	4
fitted.values	double [67]	4.35 2.25 1.73 2.25 1.73 1.73 ...
assign	integer [4]	0 1 2 3

So, the equation obtained is -

Knowhow = (-0.57625*PC1)+(0.61812PC2)+(0.53466PC3)

This is not surprising as the coefficients are from the eigenvalue matrix.

7. Write the loadings matrix, structured with components as columns and variables as rows. Using the loadings matrix, try to interpret relevant business (not mathematical, not statistical) meanings for the three principal components.

Answer:

This is Loadings Matrix as obtained in R:

	PC1	PC2	PC3
knowhow	-0.9826857	0.17882561	-0.04847891
problem_solving	-0.9964850	0.04216863	0.07238469
accountability	-0.9743858	-0.22347389	-0.02513452

Showing 1 to 3 of 3 entries, 3 total columns

The Pearson correlation coefficients serve as the loading matrix in the Principal Component Analysis (PCA) transformation.

With regards to the **interpretation of PC1**, it can be seen that it has a strong negative correlation with all three of the original variables, which are knowhow, problem-solving, and accountability. This highlights the significance of PC1 as a strong indicator of the three features, and implies that job requirements are heavily reliant on an individual's level of knowhow, problem-solving abilities, and accountability.

In contrast, **PC2** has a low correlation with knowhow and problem-solving, but a moderately high correlation with accountability. This suggests that roles that are represented by PC2 may not place as much emphasis on problem-solving or technical skills, but rather focus on the individual's level of accountability.

Finally, **PC3** exhibits nearly zero correlation with all three original variables, which makes it nearly useless in terms of interpretation with regards to the original attribute space.

8. How many principal components would you retain … a) Using the Kaiser rule? b) Using the Joliffe rule? c) Using the 80% rule? d) Your own judgment?

   Answer:

   Output from R:

```
Importance of components:
                          PC1    PC2     PC3
Standard deviation     1.7053 0.2893 0.09067
Proportion of Variance 0.9694 0.0279 0.00274
Cumulative Proportion  0.9694 0.9973 1.00000
```

```
> PCA.RATINGS$sdev^2
[1] 2.908081137 0.083697370 0.008221492
```

   a) Only PC1 has Variance > 1. So, we discard PC2 and PC3 and only retain PC1 as per Kaiser rule.
   b) Only PC1 has variance > 0.7. So, we discard PC2 and PC3 and only retain PC1 as per Joliffe rule.
   c) PC1 explains 96.94% of the variance in data. Hence we retain only PC1 and discard PC2 and PC3 as per 80% rule.
   d) Clearly, about 97% of the variance in data can be explained by just PC1. So, we can safely discard the remaining 2 features PC1 and PC2 without much loss in information.

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?

   Answer:

   The summary of the Linear Regression from R is pasted below:

```
Call:
lm(formula = salary ~ PC1 + PC2 + PC3, data = STANDARDIZED_RATINGS)

Residuals:
    Min      1Q  Median      3Q     Max
-8152.4  -865.0   189.7   628.3  6705.5

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  63929.3      254.4 251.326   <2e-16 ***
PC1          -3557.2      150.3 -23.669   <2e-16 ***
PC2          -2316.1      885.9  -2.615   0.0112 *
PC3          -3540.6     2826.5  -1.253   0.2150
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2082 on 63 degrees of freedom
Multiple R-squared:  0.9003,    Adjusted R-squared:  0.8955
F-statistic: 189.5 on 3 and 63 DF,  p-value: < 2.2e-16
```

Clearly, the **R-Squared is 90.03%** and **Adjusted R-squared is 89.55%.** So, the amount of variance in salary explained by the **three PCs is 90.03% approximately**.

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

Answer:

a) Output from R:

```
Importance of components:
                          PC1    PC2     PC3
Standard deviation     1.7053 0.2893 0.09067
Proportion of Variance 0.9694 0.0279 0.00274
Cumulative Proportion  0.9694 0.9973 1.00000
> PCA.RATINGS$sdev^2
[1] 2.908081137 0.083697370 0.008221492
```

It is evident that PC1 is the most valuable in explaining salary, accounting for a staggering 96.94% of the variance. PC2 and PC3, while still contributing to the explanation, fall behind as the second and third most valuable components. Nevertheless, they still play a role, albeit a lesser one, in explaining the variance in salary. But this role is minimal.

b) Yes the usefulness is in the order PC1>PC2>PC3. This has been explained in the above question.

c) Summary of Linear Regression from R:

```
> #Fitting Linear Regression on Salary using only PC1
> lm_salary1 = lm(salary ~ PC1, data=STANDARDIZED_RATINGS)
> summary(lm_salary1)

Call:
lm(formula = salary ~ PC1, data = STANDARDIZED_RATINGS)

Residuals:
    Min      1Q  Median      3Q     Max
-7980.5 -1028.3   340.3  1031.0  5815.5

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  63929.3      266.6  239.79   <2e-16 ***
PC1          -3557.2      157.5  -22.58   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2182 on 65 degrees of freedom
Multiple R-squared:  0.887,     Adjusted R-squared:  0.8852
F-statistic:   510 on 1 and 65 DF,  p-value: < 2.2e-16
```

Clearly, using only PC1 has **dropped the amount of variance explained** from about **90.03% (in Q9) to 88.7%**. This is a minor drop. Losing PC2 and PC3 does not impact the amount of variance explained by much.

Appendix:

Here is the entire R script that I used for this Homework-

```
# First, read in the data and set it to a dataset RATINGS
library(readxl)
RATINGS <- read_excel("Job Ratings.xlsx")
View(RATINGS)

# Extracting Principal Components on 3 selected variables by Consulting firm
(scale=TRUE).
PCA.RATINGS <- prcomp(RATINGS[2:4], scale=TRUE)

# Names of variables in the PCA object
names(PCA.RATINGS)

#Interpretation of PCs:
PCA.RATINGS$rotation

# Relative importance and summary of the PCs
summary(PCA.RATINGS)

# View data in terms of PCs
PCA.RATINGS$x
```

```
# Make data copyable to Excel
View(PCA.RATINGS$x)

# Eigenvalues of the data
m = cor(RATINGS[2:4])
eigen(m)

# Read in Standardized Original Data from Excel
STANDARDIZED_RATINGS <- read_excel("Standardized Data.xlsx")
View(STANDARDIZED_RATINGS)

# Run No Intercept Regression of three original variables (standardized) on PC1
lmPC1 = lm(PC1 ~ knowhow + problem_solving + accountability,
data=STANDARDIZED_RATINGS)

# Run No Intercept Regression of knowhow variable on PC1,PC2,PC3
lm_knowhow = lm(knowhow ~ PC1 + PC2 + PC3, data=STANDARDIZED_RATINGS)

# Correlation Matrix for Loadings
View(cor(STANDARDIZED_RATINGS)[5:7,2:4])

# Summary of Variance Explained
summary(PCA.RATINGS)

PCA.RATINGS$sdev^2

# Fitting Linear Regression on Salary using three PCs
lm_salary = lm(salary ~ PC1 + PC2 + PC3, data=STANDARDIZED_RATINGS)

#Fitting Linear Regression on Salary using only PC1
lm_salary1 = lm(salary ~ PC1, data=STANDARDIZED_RATINGS)
summary(lm_salary1)
```