MSDS 6372 : Principal Component Analysis

***Glass Identification : Refractive Index (RI) Analysis***

**1. Introduction**

“*Now you see it, now you don't. Glass is a bit of a riddle. It's hard enough to protect us, but it shatters with incredible ease. It's made from opaque sand, yet it's completely transparent. And, perhaps most surprisingly of all, it behaves like a solid material... but it's also a sort of weird liquid in disguise! You can find glass wherever you look: most rooms in your home will have a glass window and, if not that, perhaps a glass mirror... or a glass lightbulb. Glass is one of the world's oldest and most versatile human-created materials.”* Chris Woodford September 17, 2016, <http://www.explainthatstuff.com/glass.html>

Glass is made from a liquid form of sand with a temperature of . When the sand cools, it completely transforms into a different structure that is not quite solid. It becomes a material that scientists refer to as an **amorphous solid**, a cross between a solid and a liquid. Usually other materials are added into the liquid to change its chemical properties based on the purpose of the glass.

Refractive index (RI) determines how much light is bent or refracted when entering an object or in our case glass. The purpose of this study is to understand the different chemical compositions of glass and how these chemicals can affect the refractive index.

**2. Exploratory Analysis**

http://archive.ics.uci.edu/ml/datasets/Glass+Identification

|  |  |
| --- | --- |
| ID | ID number: 1 to 214 |
| RI | Refractive index |
| NA20 | Sodium (unit measurement: weight percent in corresponding oxide, as are attributes 4-10) |
| MG0 | Magnesium |
| AL203 | Aluminum |
| SI02 | Silicon |
| K20 | Potassium |
| CA0 | Calcium |
| BA0 | Barium |
| FE203 | Iron |
| Type | Type of glass |

**3. Principal Component Analysis (PCA)**

**Correlations**:

First we will take a look at the correlation matrix between the elements of sand. In our study, we are using the correlation matrix over the covariance matrix because of the variance in the variables of these data. The correlation matrix is a sums of squares and cross products from the standardized data. This correlation will tell use which elements have the highest positive and highest negative correlation.

In the correlation matrix **(Ref: Table A)**, we see a high positive correlation between variables ***Sodium Oxide***(NA20) and ***Barium Oxide***(BA0) with ***0.33***, ***Aluminum Oxide***(AL203) and ***Potassium Oxide***(K20) with **0.33** and ***Aluminum Oxide*** (AL203) and ***Barium Oxide***(BA0) with **0.48**. We also see a low negative correlation between variables ***Magnesium Oxide***(MG0) and ***Aluminum Oxide***(AL203) with ***-0.48***, ***Magnesium Oxide***(MG0) and ***Barium*** Oxide(BA0) with ***-0.49*** and ***Magnesium Oxide***(MG0) and ***Calcium Oxide***(CA0) with ***-0.44***.

|  |
| --- |
| \\.psf\Home\Desktop\Screen Shot 2016-11-10 at 9.06.25 PM.png |

**Table A – Correlation Matrix Standardized Data**

**Eigenvalues**:

Next we are going to take a look at the Eigenvalues from the Eigenvalue Matrix below and the scree plot **(Ref: Table B and C)**. Looking at the matrix the first four (4) principal components variables have Eigenvalues greater than one (1) explaining 77% of the variance. The largest difference is between component one (1) and component two (2) with 0.60.

Our Scree Plot for these data doesn’t indicate a steep curve, followed by a bend then a flat horizontal line. Instead this Scree Plot seem to descend at an angle almost linear pattern and ???. Next to the Scree Plot in Table C is the Variance Explained plot which is a graphical view of the Eigenvalue Matrix in Table B. Between the Eigenvalue Matrix and the Variance Explained graph helps us decide the number of components.

|  |
| --- |
| \\.psf\Home\Desktop\Screen Shot 2016-11-10 at 9.07.03 PM.png  **Table B – Eigenvalue Matrix** |
| \\.psf\Home\Desktop\Screen Shot 2016-11-10 at 9.08.06 PM.png |

**Components**:

From the Eigenvalues analysis, above we will focus on the first four (4) principal components more closely. In **Graph 2** shows a matrix plot of component scores between the first four components. The histogram of each component is displayed in the diagonal element of the matrix. This histogram for the first component shows that it is skewed to the right with the fourth component slightly skewed to the left. The second the third component histogram are close to normal distribution.

|  |
| --- |
| \\.psf\Home\Desktop\Screen Shot 2016-11-11 at 11.09.18 PM.png  **Graph 1 – Component Scores Matrix** |

NA20 Prin1, 0.38

MG0 Prin2, 0.40

AL203 Prin1, 0.52

SI02 Prin4, 0.83

K20 Prin2, 0.53

CA0 Prin3 0.32

BA0 Prin1 0.54

FE203 Prin3

|  |
| --- |
| \\.psf\Home\Desktop\Screen Shot 2016-11-11 at 10.21.18 PM.png |

**5. Regression Analysis of the Principal Components**

Regression analysis of the principal components….

**6. Conclusion**

**Statistical Conclusion** :

**APPENDIX**

**SAS CODE**

|  |
| --- |
| **data** glass;  infile 'glass.csv' dlm=',' firstobs=**2** dsd;  input ID RI NA20 MG0 AL203 SI02 K20 CA0 BA0 FE203 Type;  **run**;  **proc** **print** data=glass ; **run**;  **proc** **princomp** data=glass out=a ;  var NA20 MG0 AL203 SI02 K20 CA0 BA0 FE203 ;  **run**; |

**References**

http://www.explainthatstuff.com/glass.html

\*\*\*\*\*\*\*\*\*\*\*\*