Noakhali Science Technology University

Department of Information Communication Engineering

Project-01: Git Github

Student:
Mukter Alahi
Roll: 2011041
Session: 2019-20

Assistant Professor:
K.M. Aslam Uddin

May 19, 2022

NS
TU
নোয়াখালী বিজ্ঞান ও প্রযুক্তি বিশ্ববিদ্যালয়

Contents

List of Tables

List of Figures

ABSTRACT

Principal Components Analysis (PCA) and Canonical Correlation Analysis (CCA) are among the methods used in Multivariate Data Analysis. y.

Keywords: *Principal Components Analysis; Canonical correlation analysis*

1 Introduction

Principal Component Analysis is concerned with explaining the variance-covariance structure of a set of variables through a few linear combinations of these variables and its major objectives are data reduction and interpretation (Johnson and Wichern, 2007).

2 Data Description

The data set was on the characteristics of Pulp fibers and Paper made from them which comprises of eight variables and 62 observations. Measurement on characteristic of pulp fibers and the paper made from them were taken and recorded as follows:
Paper properties: $Y_1 = x_1^{(1)}$= breaking length, $Y_2 = x_2^{(1)}$= elastics module, $Y_3 = x_3^{(1)}$= stress at failure and $Y_4 = x_4^{(1)}$= burst strength.
Pulp fibers properties: $Z_1 = x_1^{(2)}$= arithmetic fiber length, $Z_2 = x_2^{(2)}$= long fiber fraction, $Z_3 = x_3^{(2)}$= fine fiber fraction and $Z_4 = x_4^{(2)}$= zero span tensile.

3 Methodology

3.1 Exploratory Data Analysis

Summary statistics (mean and standard deviation) were computed for Paper variable and Pulp fibres separately and for the combined dataset in order to see how they behave.

3.2 Principal Components Analysis

Principal component can be defined as a linear combination of measurements with maximum variability.

3.3 Canonical Correlation Analysis

The purpose of canonical correlation analysis is to identify and quantify the associations between two sets of variables, based on the correlation between a linear combination of the variables in one set and a linear combination of the variables in another set. The idea is to determine the pair of linear combinations having largest correlation among all pairs uncorrelated to the initially selected pairs and so on (Johnson and Wichern, 2007).

3.4 Software

SAS version 9.4 was used for the statistical analysis.

4 Results and Discussion

4.1 Summary Statistics

The mean and standard deviation of the pulp fiber and paper characteristic are presented in Table 1 below. We observe the difference in the variables dimensions and thus these results suggest that we should use the correlation structure to obtain the canonical variables in the canonical correlation analysis.

Table 1: Mean and standard deviation of variables for the pulp fiber and paper characteristics.

Characteristics	**Mean**	**Standard deviation**
Paper		
Breaking length (BL)	21.7228	2.8815
Elastic modulus (EM)	7.2662	0.7165
Stress at failure (SF)	5.6375	1.4629
Burst strength (BS)	1.0188	0.6930
Pulp fiber		
Arithmetic fiber length	-0.0218	0.2495
Long fiber fraction	39.0327	14.8678
Fine fiber fraction	26.6777	17.5613
Zero span tensile	1.0668	0.0295

4.2 Principal Component Analysis

To summarize the data concisely, we conducted the principal component analysis and results as seen below.

Table 2: Covariance Matrix

	Covariance Matrix			
	BL	**EM**	**SF**	**BS**
BL	8.302870935	1.88664	4.147318117	1.97206
EM	1.886636297	0.51336	0.987585105	0.43431
SF	4.147318117	0.98759	2.140045761	0.98797
BS	1.972056208	0.43431	0.987966296	0.48027

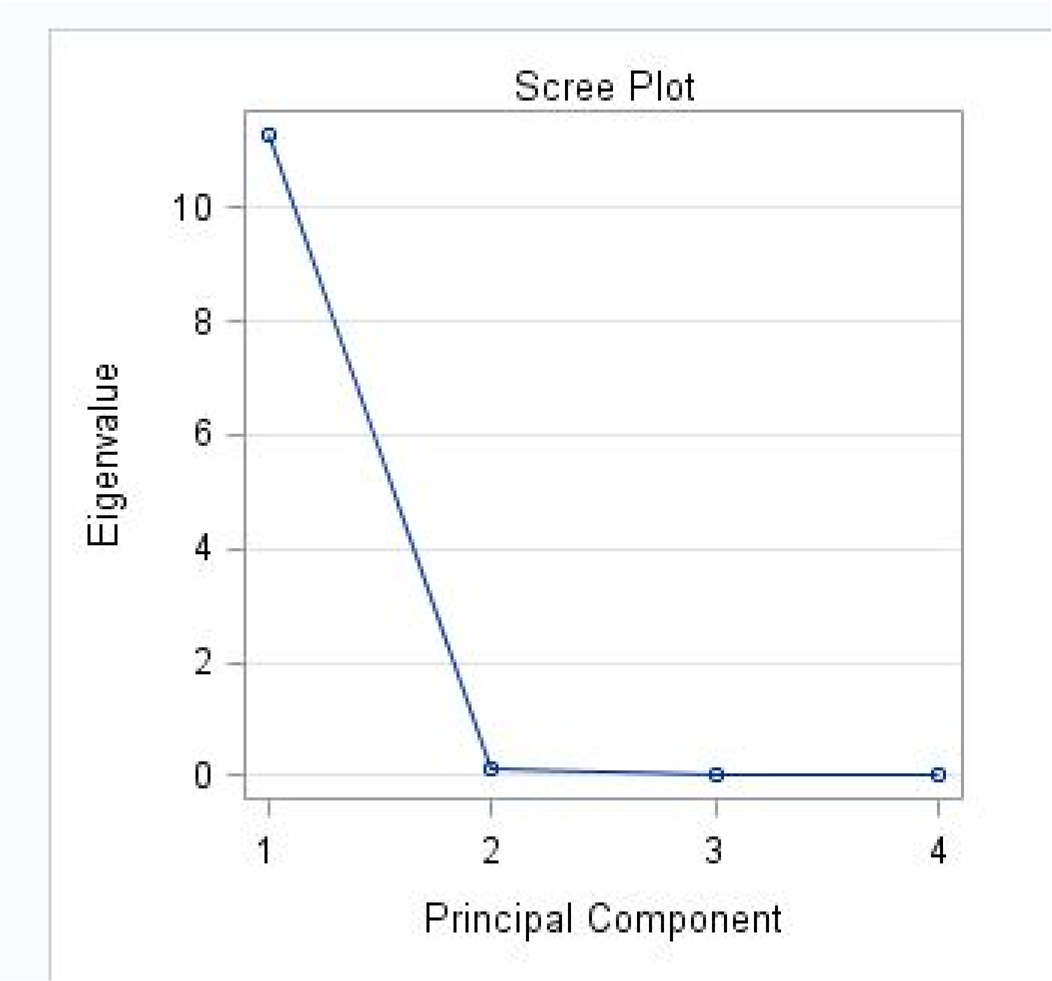
This yields a total variance of 11.436548105. The eigen values of the correlation matrix and the proportion of the variance explained by the principal components is seen in table 3 below.

From table 3 above the first eigenvalue explains 98.8% of the total variability. Thus we can summarize our four original outcome variables using only one principal component since it captures nearly all the variability explained by our four original variables.

Table 3: Eigenvalues of the Covariance Matrix

Eigenvalues of the Covariance Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
PC1	11.2950086	11.1914	0.9876	0.9876
PC2	0.1036205	0.07175	0.0091	0.9967
PC3	0.0318692	0.02582	0.0028	0.9995
PC4	0.0060497		0.0005	1

This could also be seen clearly on the scree plot below:



The eigen vectors are seen in table 4 below:

Table 4: The Eigenvectors

	Eigenvectors			
	Prin1	**Prin2**	**Prin3**	**Prin4**
BL	0.856478	-0.36392	-0.331541	-0.1552
EM	0.197573	0.78586	-0.497312	0.30995
SF	0.431271	0.45768	0.733054	-0.2592
BS	0.20351	-0.20127	0.324645	0.90149

From the eigenvectors, the first principal component can be written as

PC1=0.856(BL) + 0.198(Em) + 0.431(SF) + 0.204(BS)

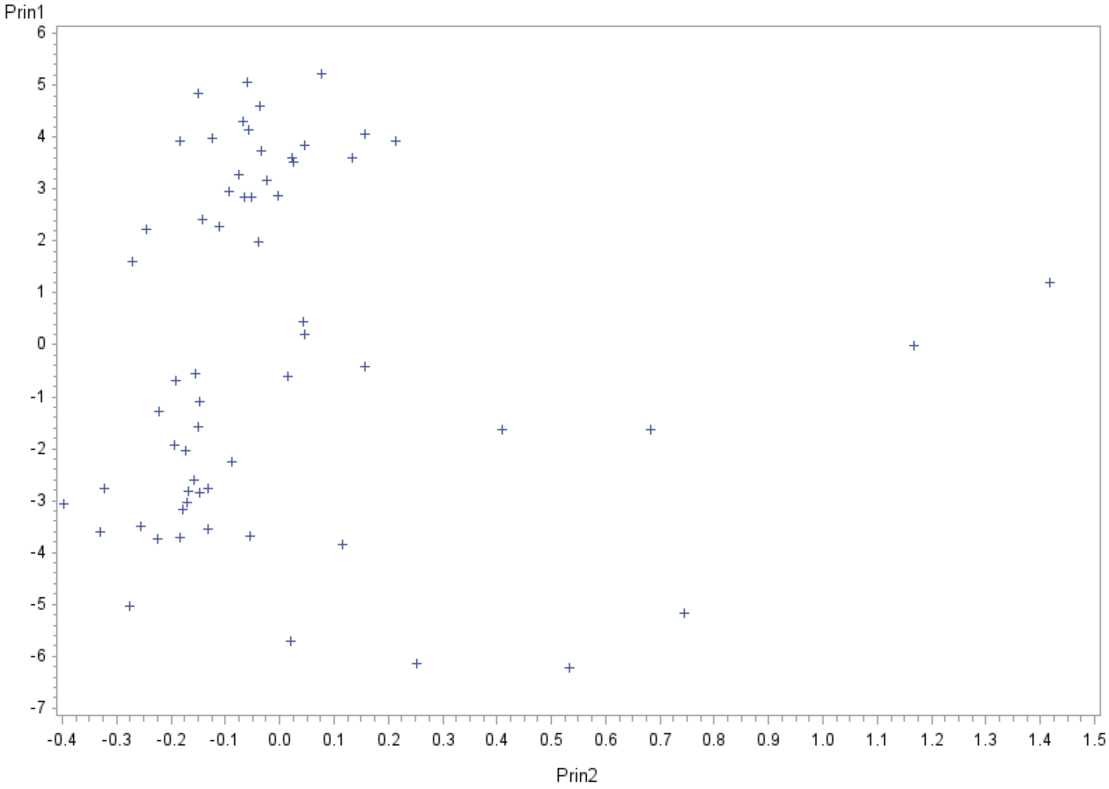
which is basically a linear combination of all the four orignal paper variables. So, it would be logical to develop a "paper strength" index for this first principal component since it does not only contain a linear combination of all the original paper variables but also captures the highest proportion of the variability explain by the paper original variables. The strength of the first principal component and the original variables can also be shown on table 5 below:

Table 5: Correlations of PC1 and PC2 with Original Paper variables

	Correlation Coefficients					
	PC1	**PC2**	**BL**	**EM**	**SF**	**BS**
PC1	1	0	0.99895	0.92675	0.99079	0.98693
PC2	0	1	-0.04066	0.35307	0.10071	-0.09349

As seen in table 5 above, there are very high correlations between the first principal component and the original variables as expected compared to the second principal component if truely the first is the best linear combination of the original variables with the highest variance.

In order to find out if there were outliers, the plot of PC1 and PC2 was performed as seen below:



As clearly seen in the graph above; there are few outlying observations in the data set (keep right on the graph)

5 Conclusion

In conclusion, from the principal component analysis, it seems the paper features can best be studies using just a single linear combination (first principal component) which explains nearly all the variability (98.7%) explained by the original paper variables. Also, from the Canonical Correlation analysis, it is observed that the first two canonical variates are good summary measures of the two sets of pulp fiber and paper variables. We were able to reduce the number of original variables from 4 to 2 pairs of canonical variables, this was due to the high canonical correlation values obtained, and the significance of the likelihood ratio statistics for the two canonical correlations. The first two canonical correlations explains 98.86% of the between groups correlations.

References

1. Green, P. E. (1978). *Analyzing Multivariate Data.* Hinsdale, Ill.: Holt, Rinehart, and Winston.
2. Johnson, R., Wichern, D. (2007). *Applied Multivariate Statistical Analysis.* 6th Ed. London: Pearson.
3. Gnanadesikan, R. (1977), *Methods for Statistical Data Analysis of Multivariate Observations, New York: John Wiley & Sons, Inc.*

APPENDIX

```
SAS CODE:
libname MDA "C:\Users\maurice\Documents\MDA\HW2";run;
data MDA.pulp;
infile 'C:\Users\maurice\Documents\MDA\HW2\pulp.txt' delimiter='09'x
MISSOVER lrecl=32767;
input  Y1 Y2 Y3 Y4 Z1 Z2 Z3 Z4;
label Y1="Breaking length"
Y2="Elastic modulus"
Y3="Stress at failure"
Y4="Burst Strength"
Z1="Arithmetic fiber length"
Z2="Long fiber fraction"
Z3="Fine fiber fraction"
Z4="Zero span tensile";
run;
proc print data=MDA.pulp; run;


/*correlation matrix**/
ods rtf ;
ods graphics on;
proc princomp data=MDA.pulp plot=matrix out=pulppcr;
var Y1 Y2 Y3 Y4;
run;

ods graphics on;
proc gplot data=pulppcr;
plot prin1*prin2;
run; quit;

######Canonical correlation Analysis ######
/*CANONICAL*/
proc cancorr data=MDA.pulp all out=pulpcc vprefix=pulp wprefix=paper
vname="Pulp Vars" wname="Paper Vars";
var Z1 Z2 Z3 Z4;
with Y1 Y2 Y3 Y4;
run;
```