Hasselt University

Master of Statistics

multivariate data analysis

Homework 2: Principial components analysis and Canonical correlation analysis

Students:
Armel Maurice Cheugoua Zanetsie
Anthony Agyapong Adomah
Mbukam Edward Chongsi
Melvis Emade Ngeme-Ndie

Lecturer:
Prof. dr. Christel Faes

December 18, 2015

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. PCA is concerned with explaining the variance-covariance structure of a set of variables through a few linear combinations of these variables. Its general objectives are data reduction and interpretation. CCA seeks to identify and quantify the associations between two sets of variables i.e Pulp fibres and Paper variables.PCA shows that the first PC already exceeds 90% of the total variability. According to the proportion of variability explained by each canonical variable , the results suggest that the first two canonical correlations seem to be sufficient to explain the structure between Pulp and Paper characteristics with 98.86%. Despite the fact that the first the two canonical variables keep 98% of common variability, 78% was kept in the pulp fiber set and about 94% of the paper set as a whole. In the proportion of opposite canonical variable,there were approximately 64% for the paper set of variables and 78% for the pulp fiber set of variables kept for the two respectively.

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). An analysis of principal components often reveals relationships that were not previously suspected and thereby allows interpretations that would not ordinarily previously suspected and thereby allows interpretations that would not ordinarily result. Principal component analysis can also be used for exploring polynomial relationships and for multivariate outlier detection (Gnanadesikan 1977).

Canonical Correlation Analysis is a multivariate statistical model that facilitates the study of inter-relationships among sets of multiple dependent variables and multiple independent variables (Green, 1978). It focuses on the correlation between a linear combination of the variables in one set and a linear combination of variables in another set. The pairs of linear combinations are called canonical variables and their correlations are called canonical correlations (Johnson and Wichern, 2007). Canonical correlation analysis determines a set of canonical variates that are orthogonal linear combinations of the variables within each set that best explain the variability both within and between sets.

In this study the main aim was to use the four paper variables (breaking length, elastic modulus, stress at failure and burst length) for the principal components analysis and use the canonical correlation analysis to study the association between these paper variables and the pulp fiber characteristics.

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. Principal component analysis looks for a few linear combinations of the variables

that can be used to summarize the data without losing too much information in the process. Here, PCA was applied for the Paper variable and Pulp fibres datasets separately.
The first principal component is calculated in such that it accounts for the greatest possible variance in the dataset. Of course, the variance of first linear combination Y_1 was as large as possible by choosing largest eigenvalue and its corresponding eigenvectors. The second principal component is calculated in the same way, with the condition that it is uncorrelated with (i.e., perpendicular to) the first principal component and that it accounts for the next highest variance. This continues until a total of p principal components have been calculated, equal to the original number of variables(Johnson and Wichern 2007).Statistics results and graphical plots are employed to summarize the result of these analyses.

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).

Here, canonical correlation analysis on the data was performed for both Paper variable and Pulp fibres datasets. The main interest is to understand the correlation between the groups.Therefore, the two linear combinations are of interest, such that their correlation is maximal (Khattree and Naik, 2000).

The root statistics (Wilks' Lambda, Pillai's Trace, Hotelling-Lawley Trace and Roy's Greatest Root) were used to test whether the canonical correlations were important. Finally, canonical redundancy analysis was conducted in order to measure the amount of variability explained by the canonical variables.

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.

4.2 Principal Component Analysis

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

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

Table 2: Covariance Matrix

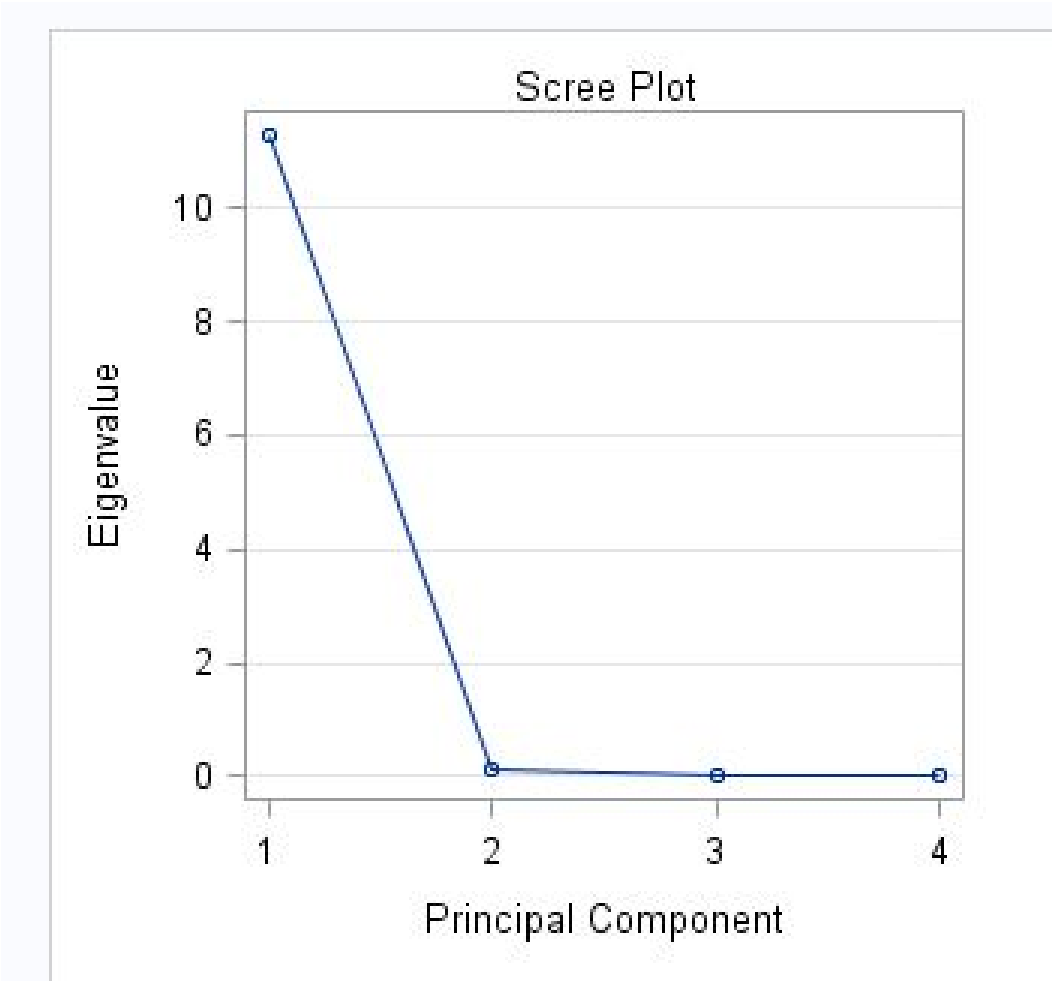
Covariance Matrix				
	BL	**EM**	**SF**	**BS**
BL	8.302870935	1.88664	4.147318117	1.97206
EM	1.886636297	0.51336	0.987585105	0.43431
SF	4.147318117	0.98759	2.140045761	0.98797
BS	1.972056208	0.43431	0.987966296	0.48027

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.

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

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.
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

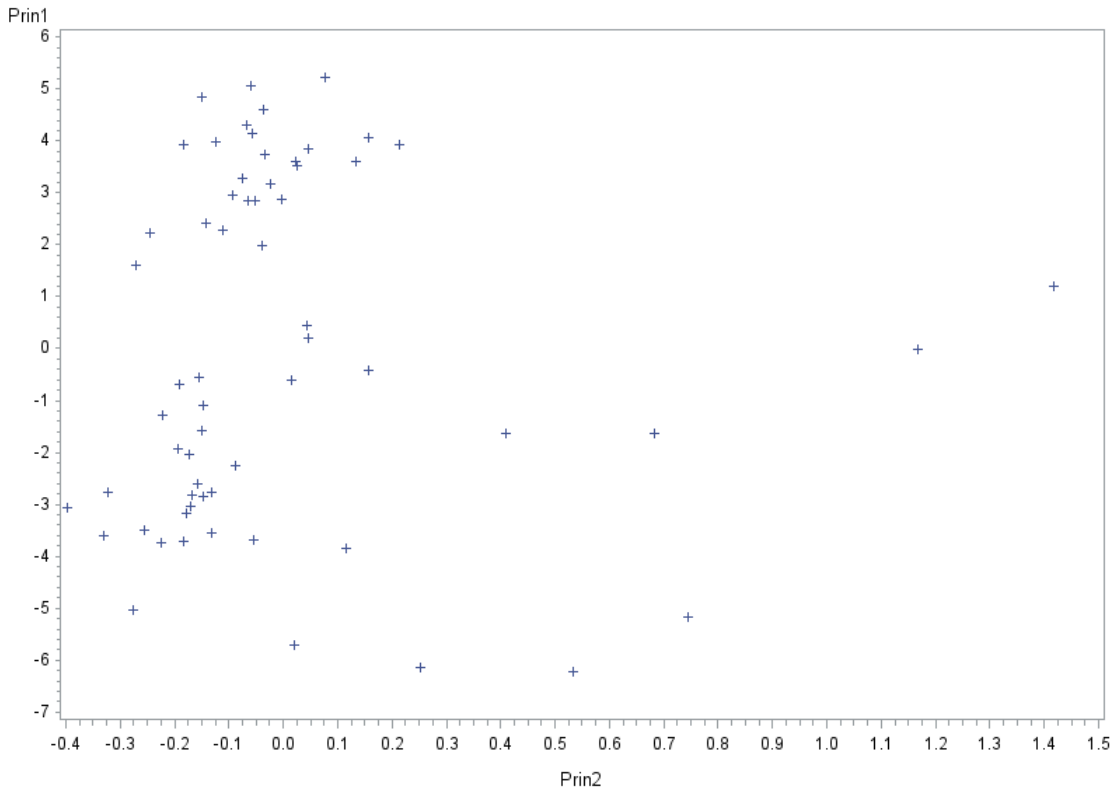
PC1=0.856(BL) + 0.198(Em) + 0.431(SF) + 0.204(BS)

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)

4.3 Canonical correlation analysis

In Table 6 below, the original correlation matrix among the variables for the Paper and Pulp fibers characteristic variables is presented. We can observe a high correlation of variables within group characteristics. The correlation between the two groups of variables was slightly lower, but still high, suggesting that the use of canonical variables maybe useful. Negative correlations between fine fiber fraction $x_3^{(2)}$ (pulp fiber characteristics) and all paper characteristics were observed, ranging from -0.57 to -0.54. For all other variables, the correlations were positive, ranging from 0.53 to 0.86 with the highest correlation being between zero span tensile $x_4^{(2)}$ and stress at failure $x_3^{(1)}$ with a value of 0.8651.

Table 6: Original Correlation Matrix of the Pulp and Paper Characteristics.

	$x_1^{(1)}$	$x_2^{(1)}$	$x_3^{(1)}$	$x_4^{(1)}$	$x_1^{(2)}$	$x_2^{(2)}$	$x_3^{(2)}$	$x_4^{(2)}$
$x_1^{(1)}$	1.0000	0.9138	0.9839	0.9876	0.6478	0.7350	-0.5419	0.8218
$x_2^{(1)}$	0.9138	1.0000	0.9422	0.8747	0.5370	0.6085	-0.5560	0.8496
$x_3^{(1)}$	0.9839	0.9422	1.0000	0.9745	0.6807	0.7644	-0.5746	0.8651
$x_4^{(1)}$	0.9876	0.8747	0.9745	1.0000	0.7064	0.7963	-0.5634	0.8132
$x_1^{(2)}$	0.6478	0.5370	0.6807	0.7064	1.0000	0.9056	-0.7334	0.7842
$x_2^{(2)}$	0.7350	0.6085	0.7644	0.7963	0.9056	1.0000	-0.7110	0.7927
$x_3^{(2)}$	-0.5419	-0.5560	-0.5746	-0.5637	-0.7334	-0.7110	1.0000	-0.7846
$x_4^{(2)}$	0.8218	0.8496	0.8651	0.8132	0.7842	0.7927	-0.7846	1.0000

The sample canonical variates are as below using the standardized canonical coefficients

$Paper1 = -1.5054x_1^{(1)} - 0.2119x_2^{(1)} + 1.9984x_3^{(1)} + 0.6764x_4^{(1)}$

$Pulp1 = -0.1593x_1^{(2)} + 0.6325x_2^{(2)} + 0.3249x_3^{(2)} + 0.8179x_4^{(2)}$

$Paper2 = -3.4956x_1^{(1)} - 1.5431x_2^{(1)} + 1.0760x_3^{(1)} + 3.7679x_4^{(1)}$

$Pulp2 = 0.6886x_1^{(2)} + 1.0029x_2^{(2)} + 0.0050x_3^{(2)} - 1.5619x_4^{(2)}$

$Paper3 = -5.7015x_1^{(1)} + 3.5252x_2^{(1)} - 4.7135x_3^{(1)} + 7.1532x_4^{(1)}$

$Pulp3 = -0.5130x_1^{(2)} + 0.0772x_2^{(2)} - 1.6631x_3^{(2)} - 0.7786x_4^{(2)}$

$Paper4 = -5.0848x_1^{(1)} - 0.5867x_2^{(1)} + 6.0694x_3^{(1)} - 0.6861x_4^{(1)}$

$Pulp4 = 2.3330x_1^{(2)} - 2.1803x_2^{(2)} + 0.0222x_3^{(2)} + 0.0891x_4^{(2)}$

From the canonical structure in table 7 below, we see that from the pulp characteristics only the variable of fine fiber fraction $x_3^{(2)}$ gives negative correlation to the covariates pairs and that all the variables in the paper characteristics gives higher correlation to the covariate pairs. The correlation between the original variables and canonical variables of one group, e.g. Paper set, are much larger than the corresponding canonical correlation.
The canonical correlations of variables 1 and 2 are high as compared to variables 3 and 4 (Table 8). This suggests that the last two canonical variables are probably not necessary.

Table 7: Correlations between the original and the canonical variables of the Standardized Paper and pulp fiber Characteristics.

	$x_1^{(1)}$	$x_2^{(1)}$	$x_3^{(1)}$	$x_4^{(1)}$	$x_1^{(2)}$	$x_2^{(2)}$	$x_3^{(2)}$	$x_4^{(2)}$
paper1	0.9351	0.8869	0.9767	0.9518	0.7491	0.8307	-0.5959	0.8618
paper2	-0.1261	-0.4280	-0.1453	0.0147	0.3009	0.3144	0.0100	-0.1885
paper3	-0.0534	0.1306	-0.0307	0.0127	0.0441	0.0472	-0.1940	0.0491
paper4	-0.3270	-0.1148	-0.1549	-0.3061	0.0378	-0.0012	-0.0191	0.0159
pulp1	0.8578	0.8136	0.8960	0.8731	0.8166	0.9056	-0.6496	0.9395
pulp2	-0.1030	-0.3496	-0.1187	0.0120	0.3683	0.3848	0.0123	-0.2307
pulp3	-0.0142	0.0346	-0.0081	0.0034	0.1661	0.1779	-0.7309	0.1851
pulp4	-0.0300	-0.0105	-0.0142	-0.0281	0.4122	-0.0126	-0.2087	0.1730

Also, the first canonical correlation is 0.9173 which is higher than any of the eight correlations between pulp fiber and paper variables (highest correlation=0.87 in table 3), indicating that we maximized the correlation between the two groups of variables. Given the magnitude of the first and second canonical correlation, one should expect the first canonical variable to explain much of the variability between sets variability but it is important to investigate whether or not the second, third and fourth canonical correlation represent much of the information contained in the data. According to the proportion of variability explained by each canonical variable in table 8, the results suggest that the first two canonical correlations seem to be sufficient to explain the structure between Pulp and Paper characteristics with 98.86%.

Table 8: Canonical Correlation, corresponding eigenvalues and Likelihood ratio test.

Canonical variable	**canonical correlation R**	**Eigenvalue**	**Proportion of variability Explained**	**Cummulative**	**Likelihood Ratio**	**P-value**
1	0.9173	5.3089	0.7175	0.7175	0.0486	$< .0001$
2	0.8169	2.0063	0.2711	0.9886	0.3066	$< .0001$
3	0.2654	0.0758	0.0102	0.9989	0.9216	0.3305
4	0.0917	0.0085	0.0011	1.0000	0.9916	0.4898

Furthermore, in order to formally assess the importance of the first canonical correlation, roots statistics comparing the within and between variability are presented in table 5. These tests are testing the null hypothesis that all canonical correlations equal zero (that is, there is no correlation between Pulp fibers and Paper characteristics sets). The four root statistics were found to be highly significant at 5% level of significance, indicating that there is an important correlation between the pulp fiber and paper variables. In order to assess how many canonical correlations are non-zero correlations, the likelihood ratio test for each canonical variable is presented in Table 8. This result indicates that the canonical correlations of the canonical variables 1 and 2 are different from zero, as previously suggested by the proportion of variability these two variables explain. Thus implying that there is a common structure described by two canonical variables.

Table 9: Multivariate Statistics and F Approximations

Statistics	Value	F-value	P-value
Wilks' Lambda	0.045	17.50	$< .0001$
Pillai's Trace	1.653	9.38	$< .0001$
Hotelling-Lawley Trace	7.324	24.53	$< .0001$
Roy's Greatest Root	5.309	75.65	$< .0001$

From the canonical redundancy analysis with standardized variance, although the first two canonical variables keep 98% of common variability, they keep only about 78% in the pulp fiber set and about 94% of the paper set as a whole. Also, the common structure is approximately 64% for the paper set of variables and 78% for the pulp fiber set of variables . This can be seen in table 10 below

Table 10: Canonical redundancy analysis with standardized variance.

	Pulp fiber variables		Paper variables	
Canonical variable	Proportion of the own canonical variable	Proportion of the opposite canonical variable	Proportion of the own canonical variable	Proportion of the opposite canonical variable
1	0.6979	0.5873	0.8802	0.7407
2	0.0843	0.0562	0.0551	0.0368
3	0.1570	0.0111	0.0052	0.0004
4	0.0609	0.0005	0.0594	0.0005

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;

######Principal component analysis####
/*covariance matrix**/
ods rtf ;
ods graphics on;
proc princomp data=MDA.pulp cov plot=matrix out=pulppc;
var Y1 Y2 Y3 Y4;
run;

ods graphics off;
ods rtf close;
proc gplot data=pulppc;
plot prin1*prin2;
plot prin2*prin3;
plot prin3*prin4;
plot prin1*prin3;
plot prin1*prin4;
run;quit;

/*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;
```