Multivariate Analysis
of
Chemical Compounds in Marijuana Data,
Socio Economic Data,
and
Genuine and Fake Bank Notes Data.

Tarun.S.Sarode

This is a Report where the concepts and techniques of Multivariate analysis is applied to gain Insight on different scenario and data corresponding to them where certain questions will be analysed and answered

Questions that will be answered in this report by applying the following techniques:

- Principal Component Analysis (PCA).
- Factor Analysis (FA) **.**
- Determinant Analysis (DA).

Principal Component Analysis.

The data is a concentration of 13 different chemical compounds in marijuana plants own in the same region in Colombia that are derived from three different species varieties. The analysis is carries on using **SAS.**

1) Compute the mean and standard deviation for the 13 chemical concentrations on the sample data via SAS

2) Compute the correlation matrix and a scatterplot in SAS. Is the correlation matrix suitable for a principal component analysis?

3) Perform a Principal component analysis using SAS on the raw data and assess how many PCs need to retain.
   a) What percentage of the total sample variation is accounted for the first, second and third PCs?
   b) Interpret the first 3 PC's.
   c) What are the first, second and third PCs as linear functions of the original variables.
   d) Can the data be effectively summarised in fewer than 13 dimensions?
   e) Visualise the number of PCs considered.

4) Perform a principal component analysis on the correlation matrix.
   a) What percentage of the total sample variation is accounted for the first, second and third PCs?
   b) Interpret the first 3 PC's.
   c) What are the first, second and third PCs as linear functions of the standardised variables.
   d) Can the data be effectively summarized in fewer than 13 dimensions?
   e) Visualise the number of PCs considered.

Solutions:

1) Computing the mean and standard deviation for the 13 chemical concentrations in the sample data.

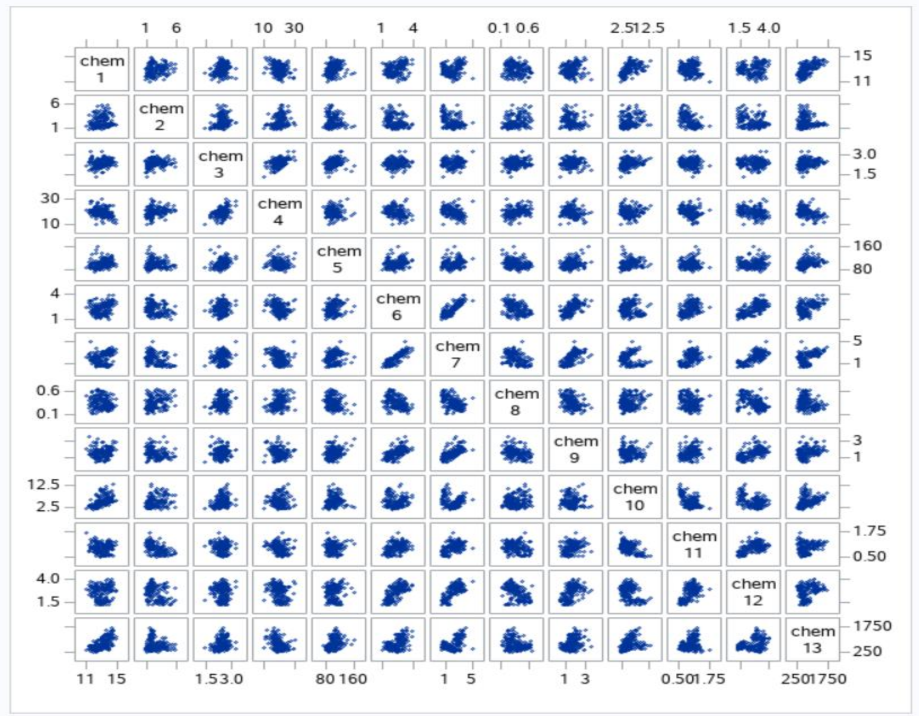
The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
varieties	178	1.9382022	0.7750350	1.0000000	3.0000000
chem1	178	13.0006180	0.8118265	11.0300000	14.8300000
chem2	178	2.3363483	1.1171461	0.7400000	5.8000000
chem3	178	2.3665169	0.2743440	1.3600000	3.2300000
chem4	178	19.4949438	3.3395638	10.6000000	30.0000000
chem5	178	99.7415730	14.2824835	70.0000000	162.0000000
chem6	178	2.2951124	0.6258510	0.9800000	3.8800000
chem7	178	2.0292697	0.9988587	0.3400000	5.0800000
chem8	178	0.3618539	0.1244533	0.1300000	0.6600000
chem9	178	1.5908989	0.5723589	0.4100000	3.5800000
chem10	178	5.0580899	2.3182859	1.2800000	13.0000000
chem11	178	0.9574494	0.2285716	0.4800000	1.7100000
chem12	178	2.6116854	0.7099904	1.2700000	4.0000000
chem13	178	746.8932584	314.9074743	278.0000000	1680.00

2) Computing the correlation matrix and a scatterplot for analysis.

Correlation Matrix

	chem1	chem2	chem3	chem4	chem5	chem6	chem7	chem8	chem9	chem10	chem11	chem12	chem13
chem1	1.0000	0.0944	0.2115	-.3102	0.2708	0.2891	0.2368	-.1559	0.1367	0.5464	-.0717	0.0723	0.6437
chem2	0.0944	1.0000	0.1640	0.2885	-.0546	-.3352	-.4110	0.2930	-.2207	0.2490	-.5613	-.3687	-.1920
chem3	0.2115	0.1640	1.0000	0.4434	0.2866	0.1290	0.1151	0.1862	0.0097	0.2589	-.0747	0.0039	0.2236
chem4	-.3102	0.2885	0.4434	1.0000	-.0833	-.3211	-.3514	0.3619	-.1973	0.0187	-.2740	-.2768	-.4406
chem5	0.2708	-.0546	0.2866	-.0833	1.0000	0.2144	0.1958	-.2563	0.2364	0.2000	0.0554	0.0660	0.3934
chem6	0.2891	-.3352	0.1290	-.3211	0.2144	1.0000	0.8646	-.4499	0.6124	-.0551	0.4337	0.6999	0.4981
chem7	0.2368	-.4110	0.1151	-.3514	0.1958	0.8646	1.0000	-.5379	0.6527	-.1724	0.5435	0.7872	0.4942
chem8	-.1559	0.2930	0.1862	0.3619	-.2563	-.4499	-.5379	1.0000	-.3658	0.1391	-.2626	-.5033	-.3114
chem9	0.1367	-.2207	0.0097	-.1973	0.2364	0.6124	0.6527	-.3658	1.0000	-.0252	0.2955	0.5191	0.3304
chem10	0.5464	0.2490	0.2589	0.0187	0.2000	-.0551	-.1724	0.1391	-.0252	1.0000	-.5218	-.4288	0.3161
chem11	-.0717	-.5613	-.0747	-.2740	0.0554	0.4337	0.5435	-.2626	0.2955	-.5218	1.0000	0.5655	0.2362
chem12	0.0723	-.3687	0.0039	-.2768	0.0660	0.6999	0.7872	-.5033	0.5191	-.4288	0.5655	1.0000	0.3128
chem13	0.6437	-.1920	0.2236	-.4406	0.3934	0.4981	0.4942	-.3114	0.3304	0.3161	0.2362	0.3128	1.0000



As can be seen from the Correlation Matrix and the Scatter plot, some of the chemicals have a strong correlation (greater than 0.44) with each other

Namely;
- chem6&chem7
- chem7&chem9
- chem1&chem10
- chem7&chem12
- chem11&chem12 and
- chem1&chem13

3) Perform a Principal component analysis using SAS on the raw data and assess how many PCs need to retain.

a) What percentage of the total sample variation is accounted for the first, second and third PCs?

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	99202.0307	99029.4941	0.9981	0.9981
2	172.5366	163.0054	0.0017	0.9998
3	9.5312	4.4304	0.0001	0.9999
4	5.1008	3.8150	0.0001	1.0000
5	1.2858	0.4176	0.0000	1.0000
6	0.8682	0.5812	0.0000	1.0000
7	0.2870	0.1317	0.0000	1.0000
8	0.1553	0.0415	0.0000	1.0000
9	0.1137	0.0274	0.0000	1.0000
10	0.0864	0.0402	0.0000	1.0000
11	0.0462	0.0113	0.0000	1.0000
12	0.0349	0.0142	0.0000	1.0000
13	0.0208	0.0127	0.0000	1.0000
14	0.0081		0.0000	1.0000

The percentage of total sample variation are:

1st = 99.8%

2nd = 0.17%

3rd = 0.1%

b) Interpret the first 3 PC's.

As from the PC's we can see that only first PC which is 99.8% is high and significant whereas the second and third are very less significant which would not be of significance if included.

c) Write out the first, second and third PCs as linear functions of the original variables.

Eigenvectors														
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12	Prin13	Prin14
varieties	-.001559	0.002779	0.100920	0.147646	-.219197	-.168685	-.178801	0.190995	0.165828	0.510993	0.623529	0.366354	0.095988	0.047158
chem1	0.001659	0.001204	0.018304	0.136989	0.048933	0.202080	0.890116	0.349028	0.104297	0.057590	0.089075	0.034777	-.004240	0.013422
chem2	-.000681	0.002160	0.124559	0.156739	-.528931	0.803840	-.146858	-.075521	0.010226	-.033610	-.007744	-.035736	0.063236	-.012055
chem3	0.000195	0.004594	0.051293	-.012047	0.025846	0.041694	0.050349	-.150528	0.068881	-.107161	-.353256	0.891487	-.086488	-.162345
chem4	-.004671	0.026461	0.928100	-.356990	0.069611	-.023860	0.035937	0.011392	-.002441	-.009980	0.002659	-.059311	0.000269	-.000065
chem5	0.017868	0.999340	-.029985	-.004754	-.006452	-.001349	0.002065	-.003569	-.001632	0.000865	0.001638	-.002620	0.000501	0.002271
chem6	0.000990	0.000875	-.042761	-.076452	0.320081	0.228320	-.065680	-.089903	0.364232	0.719696	-.393379	-.135677	-.019082	-.034648
chem7	0.001567	-.000059	-.090267	-.172191	0.535696	0.357645	-.079455	-.204900	0.384930	-.318459	0.485222	0.080513	-.000197	0.086672
chem8	-.000123	-.001354	0.013722	0.010594	-.029289	-.016952	-.000054	0.000990	0.028873	-.017794	-.175069	0.116465	0.132954	0.967218
chem9	0.000601	0.005002	-.026237	-.051606	0.253787	0.197181	-.351740	0.847728	-.121294	-.104024	-.131390	0.087044	-.015025	-.019293
chem10	0.002327	0.015114	0.303203	0.856516	0.367299	-.005508	-.093290	-.108372	-.110801	-.044505	-.035782	-.037635	0.042567	-.007837
chem11	0.000171	-.000764	-.026992	-.059055	0.045943	-.030413	0.029583	-.002817	0.027086	-.051943	-.082710	0.024792	0.978655	-.152355
chem12	0.000705	-.003501	-.074366	-.178526	0.269297	0.249151	0.079984	-.180924	-.804595	0.296076	0.179432	0.130239	0.038019	0.056751
chem13	0.999822	-.017769	0.004627	-.002951	-.002713	-.001211	-.001178	0.000095	0.000030	0.000388	0.000529	0.000047	-.000040	0.000067

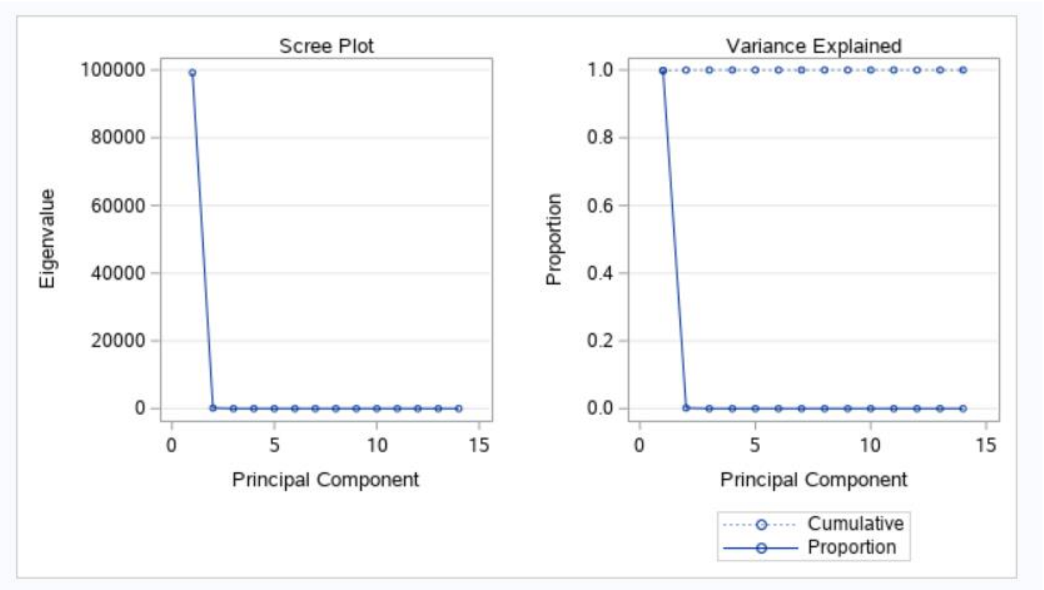
Prin1 = 0.001659* chem1 -.000681* chem2 + 0.000195* chem3 -.004671* chem4 + 0.017868* chem5 + 0.000990*chem6 + 0.001567*chem7 -.000123*chem8 + 0.000601*chem9 + 0.002327*chem10 + 0.000171*chem11 + 0.000705 *chem12 + 0.999822*chem13

Prin2 = 0.001204*chem1 + 0.002160*chem2 + 0.004594*chem3 + 0.026461*chem4 + 0.999340*chem5 + 0.000875*chem6 -.000059*chem7 -.001354*chem8 + 0.005002*chem9 + 0.015114*chem10 -.000764*chem11 -.003501*chem12 -.017769*chem13

Prin3 = 0.018304*chem1 + 0.124559*chem2 + 0.051293*chem3 + 0.928100*chem4 -.029985*chem5 -.042761*chem6 -.090267*chem7 + 0.013722*chem8 -.026237*chem9 + 0.303203*chem10 -.026992*chem11 -.074366*chem12 + 0.004627*chem13

d) Can the data be effectively summarised in fewer than 13 dimensions?
From the results we can see that as the PC1 cumulatively represents 99.8% of variance therefore the data can be summarised in fewer than 13 dimensions.

e) Visualise the number of PCs considered.



From the Screen Plot we can see that the change is at point 2 and from all there all the values are near zero which confirms that first PC can summarise the data.

4) Perform a principal component analysis using SAS on the correlation matrix. Answer the following from the resultant output

a) What percentage of the total sample variation is accounted for the first, second and third PCs?

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	4.70585025	2.20887652	0.3620	0.3620
2	2.49697373	1.05090176	0.1921	0.5541
3	1.44607197	0.52709805	0.1112	0.6653
4	0.91897392	0.06574575	0.0707	0.7360
5	0.85322818	0.21157115	0.0656	0.8016
6	0.64165703	0.09062872	0.0494	0.8510
7	0.55102831	0.20253095	0.0424	0.8934
8	0.34849736	0.05961742	0.0268	0.9202
9	0.28887994	0.03797746	0.0222	0.9424
10	0.25090248	0.02511384	0.0193	0.9617
11	0.22578864	0.05701840	0.0174	0.9791
12	0.16877023	0.06539230	0.0130	0.9920
13	0.10337794		0.0080	1.0000

The percentage of total sample variation are:

1st = 36.2%

2nd = 19.21%

3rd = 11.12%

b) Interpret the first 3 PC's.

As from the PC's we can see that only first PC which is 36.2% 99.8% the second PC is 19.21% and third is 11.12% which does not include most of the data.

c) What are the first, second and third PCs as linear functions of the standardised variables.

Eigenvectors													
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12	Prin13
chem1	0.144329	0.483652	-.207383	-.017856	0.265664	0.213539	-.056396	0.396139	0.508619	-.211605	-.225917	-.266286	0.014970
chem2	-.245188	0.224931	0.089013	0.536890	-.035214	0.536814	0.420524	0.065827	-.075283	0.309080	0.076486	0.121696	0.025964
chem3	-.002051	0.316069	0.626224	-.214176	0.143025	0.154475	-.149171	-.170260	-.307694	0.027125	-.498691	-.049622	-.141218
chem4	-.239320	-.010591	0.612080	0.060859	-.066103	-.100825	-.286969	0.427970	0.200449	-.052799	0.479314	-.055743	0.091683
chem5	0.141992	0.299634	0.130757	-.351797	-.727049	0.038144	0.322883	-.156361	0.271403	-.067870	0.071289	0.062220	0.056774
chem6	0.394661	0.065040	0.146179	0.198068	0.149318	-.084122	-.027925	-.405934	0.286035	0.320131	0.304341	-.303882	-.463908
chem7	0.422934	-.003360	0.150682	0.152295	0.109026	-.018920	-.060685	-.187245	0.049578	0.163151	-.025694	-.042899	0.832257
chem8	-.298533	0.028779	0.170368	-.203301	0.500703	-.258594	0.595447	-.233285	0.195501	-.215535	0.116896	0.042352	0.114040
chem9	0.313429	0.039302	0.149454	0.399057	-.136860	-.533795	0.372139	0.368227	-.209145	-.134184	-.237363	-.095553	-.116917
chem10	-.088617	0.529996	-.137306	0.065926	0.076437	-.418644	-.227712	-.033797	0.056218	0.290775	0.031839	0.604222	-.011993
chem11	0.296715	-.279235	0.085222	-.427771	0.173615	0.105983	0.232076	0.436624	0.085828	0.522399	-.048212	0.259214	-.089889
chem12	0.376167	-.164496	0.166005	0.184121	0.101161	0.265851	-.044764	-.078108	0.137227	-.523706	0.046423	0.600959	-.156718
chem13	0.286752	0.364903	-.126746	-.232071	0.157869	0.119726	0.076805	0.120023	-.575786	-.162116	0.539270	-.079402	0.014447

Prin1 = 0.144329*chem1 -.245188*chem2 -.002051*chem3 -.23932 *chem4 + 0.141992*chem5 + 0.394661*chem6 + 0.422934*chem7 -.298533*chem8 + 0.313429*chem9 -.088617*chem10 + 0.296715*chem11 + 0.376167*chem12 + 0.286752*chem13

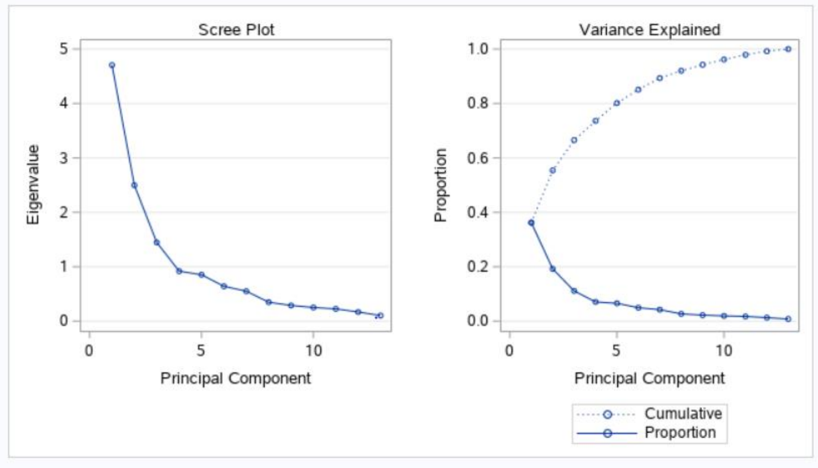
Prin2 = 0.483652*chem1 + 0.224931*chem2 + 0.316069*chem3 -.010591*chem4 + 0.299634*chem5 + 0.065040*chem6 -.00336*chem7 + 0.02877*chem8 + 0.039302 *chem9 + 0.529996*chem10 + -.279235*chem11-.164496*chem12 + 0.364903*chem13

Prin3 = -.207383*chem1 + 0.089013*chem2 + 0.626224*chem3 + 0.612080*chem4 + 0.130757*chem5 + 0.146179*chem6 + 0.150682*chem7 + 0.170368*chem8 + 0.149454*chem9 -.137306*chem10 + 0.085222*chem11 + 0.166005*chem12 -.126746*chem13

d) Can the data be effectively summarized in fewer than 13 dimensions?

From the analysis we need 5 PC to make sure that the data can be summarised and the data can be summarised in less that 13 which is 5 PC.

e) Visualise the number of PCs considered.



From the screen plot we can see that there is an elbow(bend), after which the remaining eigenvalues are relatively small or roughly the same size. The plot on the left the elbow curve occurs at PC4 and therefore it would suggest to use 4 PC's

Factor analysis

A raw data with 12 observations, on 5 socio-economic variables, called Population, School, Employment, Services and House Value will be used to carry on the analysis.

1) Compute the means and standard deviations of the data.

2) Compute a Factor analysis on the raw data and the correlation matrix.

3) From the eigenvalues of the correlation matrix and the factor loading matrix and communalities outputted the following questions can be answered.
   a) Do the first two principal components (factors) provide an adequate summary of the data?
   b) How much of the variation is accounted for by 2 factors?
   c) How much of the variation is accounted for by 3 factors?

4) Using PROC PRINCOMP to display the scoring coefficients as eigenvectors, and answer the following questions
   a) What are the eigenvalues and the respective eigenvectors?
   b) What is the proportion of the variance accounted for by the first and second component respectively?

c) Together how much do the first and second factors together account for the standardised variance?

d) Do the final communality estimates show that all the variables are well accounted for by how many components or factors?

5) To obtain the component scores as linear combinations of the observed variables along with the standardized scoring. As each factor/component can expressed as a linear combination of the standardised observed variables the following questions can be answered.

a) Write down the first principal component or Factor1 in terms of the standardised variables.

b) Write down the second principal component or Factor2 in terms of the standardised variables.

c) Write the first and second PCs in terms of eigenvectors.

Solution

1) Compute the means and standard deviations of the data.

The MEANS Procedure

Variable	Mean	Std Dev
Population	6241.6667	3439.9943
School	11.4417	1.7865
Employment	2333.3333	1241.2115
Services	120.8333	114.9275
HouseValue	17000.0000	6367.5313

2) Compute a Factor analysis on the raw data and the correlation matrix.

The FACTOR Procedure

Input Data Type	Raw Data
Number of Records Read	12
Number of Records Used	12
N for Significance Tests	12

Means and Standard Deviations from 12 Observations

Variable	Mean	Std Dev
Population	6241.667	3439.9943
School	11.442	1.7865
Employment	2333.333	1241.2115
Services	120.833	114.9275
HouseValue	17000.000	6367.5313

Correlations

	Population	School	Employment	Services	HouseValue
Population	1.00000	0.00975	0.97245	0.43887	0.02241
School	0.00975	1.00000	0.15428	0.69141	0.86307
Employment	0.97245	0.15428	1.00000	0.51472	0.12193
Services	0.43887	0.69141	0.51472	1.00000	0.77765
HouseValue	0.02241	0.86307	0.12193	0.77765	1.00000

The FACTOR Procedure
Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 5 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	2.87331359	1.07665350	0.5747	0.5747
2	1.79666009	1.58182321	0.3593	0.9340
3	0.21483689	0.11490283	0.0430	0.9770
4	0.09993405	0.08467868	0.0200	0.9969
5	0.01525537		0.0031	1.0000

2 factors will be retained by the MINEIGEN criterion.

Factor Pattern

	Factor1	Factor2
Population	0.58096	0.80642
School	0.76704	-0.54476
Employment	0.67243	0.72605
Services	0.93239	-0.10431
HouseValue	0.79116	-0.55818

Variance Explained by Each Factor

Factor1	Factor2
2.8733136	1.7966601

Final Communality Estimates: Total = 4.669974

Population	School	Employment	Services	HouseValue
0.98782629	0.88510555	0.97930583	0.88023562	0.93750041

3) From the eigenvalues of the correlation matrix and the factor loading matrix and communalities outputted the following questions can be answered.

a) Do the first two principal components (factors) provide an adequate summary of the data?
Ans: As the first two PC which explains 93.4% of variance of the dataset, we can say that it is adequate to summarise the data as values>90% is generally considered adequate.

b) How much of the variation is accounted for by 2 factors?
Ans: The first 2 factors explain 93.4% of the variance of the dataset or 4.67/5

c) How much of the variation is accounted for by 3 factors?
Ans: The first 3 factors explain 97.7% of the variance of the dataset or 4.885/5

4) Using PROC PRINCOMP to display the scoring coefficients as eigenvectors, and answer the following questions

The PRINCOMP Procedure

Observations	12
Variables	5

Simple Statistics

	Population	School	Employment	Services	HouseValue
Mean	6241.666667	11.44166667	2333.333333	120.8333333	17000.00000
StD	3439.994274	1.78654483	1241.211529	114.9275134	6367.53128

Correlation Matrix

	Population	School	Employment	Services	HouseValue
Population	1.0000	0.0098	0.9724	0.4389	0.0224
School	0.0098	1.0000	0.1543	0.6914	0.8631
Employment	0.9724	0.1543	1.0000	0.5147	0.1219
Services	0.4389	0.6914	0.5147	1.0000	0.7777
HouseValue	0.0224	0.8631	0.1219	0.7777	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.87331359	1.07665350	0.5747	0.5747
2	1.79666009	1.58182321	0.3593	0.9340
3	0.21483689	0.11490283	0.0430	0.9770
4	0.09993405	0.08467868	0.0200	0.9969
5	0.01525537		0.0031	1.0000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5
Population	0.342730	0.601629	0.059517	0.204033	0.689497
School	0.452507	-.406414	0.688822	-.353571	0.174861
Employment	0.396695	0.541665	0.247958	0.022937	-.698014
Services	0.550057	-.077817	-.664076	-.500386	-.000124
HouseValue	0.466738	-.416429	-.139649	0.763182	-.082425

a) What are the eigenvalues and the respective eigenvectors?

	Eigenvalue
1	2.87331359
2	1.79666009
3	0.21483689
4	0.09993405
5	0.01525537

Eigenvectors					
	Prin1	Prin2	Prin3	Prin4	Prin5
Population	0.342730	0.601629	0.059517	0.204033	0.689497
School	0.452507	-.406414	0.688822	-.353571	0.174861
Employment	0.396695	0.541665	0.247958	0.022937	-.698014
Services	0.550057	-.077817	-.664076	-.500386	-.000124
HouseValue	0.466738	-.416429	-.139649	0.763182	-.082425

b) What is the proportion of the variance accounted for by the first and second component respectively?

Ans:

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.87331359	1.07665350	0.5747	0.5747
2	1.79666009	1.58182321	0.3593	0.9340
3	0.21483689	0.11490283	0.0430	0.9770
4	0.09993405	0.08467868	0.0200	0.9969
5	0.01525537		0.0031	1.0000

The first and the second component account for 93.4% of the variance.

c) Together how much do the first and second factors together account for the standardised variance?

Ans: The first 2 factors explain 93.4% of the variance. This is the same as the previous question because PCA was used to perform the estimation for the factor analysis.

d) Do the final communality estimates show that all the variables are well accounted for by how many components or factors?

Ans:

Final Communality Estimates: Total = 4.669974				
Population	School	Employment	Services	HouseValue
0.98782629	0.88510555	0.97930583	0.88023562	0.93750041

The final communality estimates are between 0.987 for population and 0.8802 for service whereas the communality estimate is 4.66.

5) To obtain the component scores as linear combinations of the observed variables along with the standardized scoring. As each factor/component can expressed as a linear combination of the standardised observed variables the following questions can be answered.

The FACTOR Procedure
Initial Factor Method: Principal Components

Scoring Coefficients Estimated by Regression

Squared Multiple Correlations of the Variables with Each Factor				
Factor1	Factor2	Factor3	Factor4	Factor5
1.0000000	1.0000000	1.0000000	1.0000000	1.0000000

Standardized Scoring Coefficients					
	Factor1	Factor2	Factor3	Factor4	Factor5
Population	0.20219	0.44884	0.12841	0.64542	5.58240
School	0.26695	-0.30320	1.48612	-1.11846	1.41574
Employment	0.23403	0.40411	0.53496	0.07256	-5.65135
Services	0.32450	-0.05806	-1.43273	-1.58288	-0.00100
HouseValue	0.27535	-0.31068	-0.30129	2.41419	-0.66734

a) Write down the first principal component or Factor1 in terms of the standardised variables.

Ans: Factor 1 = 0.20219*population + 0.26695*school + 0.23403*employment + 0.32450*services +0.27535*house value.

b) Write down the second principal component or Factor2 in terms of the standardised variables.

Ans: Factor 2 = 0.44884*population – 0.30320*school +0.40411*employment – 0.05806*services – 0.31068*house value.

c) Write the first and second PCs in terms of eigenvectors.

Ans:

Eigenvectors					
	Prin1	Prin2	Prin3	Prin4	Prin5
Population	0.342730	0.601629	0.059517	0.204033	0.689497
School	0.452507	-.406414	0.688822	-.353571	0.174861
Employment	0.396695	0.541665	0.247958	0.022937	-.698014
Services	0.550057	-.077817	-.664076	-.500386	-.000124
HouseValue	0.466738	-.416429	-.139649	0.763182	-.082425

Print1 = 0.342730*population + 0.452507*school + 0.396695*employment + 0.550057*services + 0.466738*house value.

Print 2 = 0.601629*population – 0.406414*school + 0.541665*employment – 0.077817*services – 0.416429*house value.

Determinant Analysis

For Determinant Analysis a bank data consisting of Six variables measured on 100 **genuine** and 100 **forged** (counterfeit/fake) old Swiss 1000-franc bank notes are used and the following questions are answered.

1) Compute the means and the variance-covariance matrix of the data for the **genuine notes**.

2) Compute the means and standard deviations and the variance-covariance matrix of the data for the **forged/fake/counterfeit notes**.

3) Produce the correlation matrix and an associated scatterplot of the inputted data for the **genuine notes**.

4) Produce the correlation matrix and an associated Scatterplot of the inputted data for the **forged /fake notes**.

5) Run the discriminant analysis using the SAS which allocates a bank note with the following characteristics X_0^T = (214.9, 130.1, 129.9, 9, 10.6, 140.5) to the appropriate grouping i.e., allocates it to either the **genuine** or the **forged/fake class.**

6) Using the SAS DISCRIM and resultant output answer the following questions.
   a) Is $\sum_1 = \sum_2$?
   b) How is the bank note with X_0^T = (214.9, 130.1, 129.9, 9, 10.6, 140.5) allocated?

Solution:

1) Compute the means and the variance-covariance matrix of the data for the **genuine notes.**

The MEANS Procedure

Variable	Mean
Length	214.9690000
Left	129.9430000
Right	129.7200000
Bottom	8.3050000
Top	10.1680000
Diagonal	141.5170000

Covariance Matrix, DF = 99

	Length	Left	Right	Bottom	Top	Diagonal
Length	0.1502414141	0.0580131313	0.0572929293	0.0571262626	0.0144525253	0.0054818182
Left	0.0580131313	0.1325767677	0.0858989899	0.0566515152	0.0490666667	-.0430616162
Right	0.0572929293	0.0858989899	0.1262626263	0.0581818182	0.0306464646	-.0237777778
Bottom	0.0571262626	0.0566515152	0.0581818182	0.4132070707	-.2634747475	-.0001868687
Top	0.0144525253	0.0490666667	0.0306464646	-.2634747475	0.4211878788	-.0753090909
Diagonal	0.0054818182	-.0430616162	-.0237777778	-.0001868687	-.0753090909	0.1998090909

2) Compute the means and standard deviations and the variance-covariance matrix of the data for the **forged/fake/counterfeit notes**.

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
Length	100	214.8230000	0.3521521	213.9000000	216.3000000
Left	100	130.3000000	0.2550500	129.6000000	130.8000000
Right	100	130.1930000	0.2982288	129.3000000	131.1000000
Bottom	100	10.5300000	1.1319510	7.4000000	12.7000000
Top	100	11.1330000	0.6359682	9.1000000	12.3000000
Diagonal	100	139.4500000	0.5578639	137.8000000	140.6000000

Covariance Matrix, DF = 99

	Length	Left	Right	Bottom	Top	Diagonal
Length	0.124011111	0.031515152	0.024001010	-0.100595960	0.019435354	0.011565657
Left	0.031515152	0.065050505	0.046767677	-0.024040404	-0.011919192	-0.005050505
Right	0.024001010	0.046767677	0.088940404	-0.018575758	0.000132323	0.034191919
Bottom	-0.100595960	-0.024040404	-0.018575758	1.281313131	-0.490191919	0.238484848
Top	0.019435354	-0.011919192	0.000132323	-0.490191919	0.404455556	-0.022070707
Diagonal	0.011565657	-0.005050505	0.034191919	0.238484848	-0.022070707	0.311212121

3) Produce the correlation matrix and an associated scatterplot of the inputted data for the **genuine notes**.

The CORR Procedure

6 Variables:	Length Left Right Bottom Top Diagonal

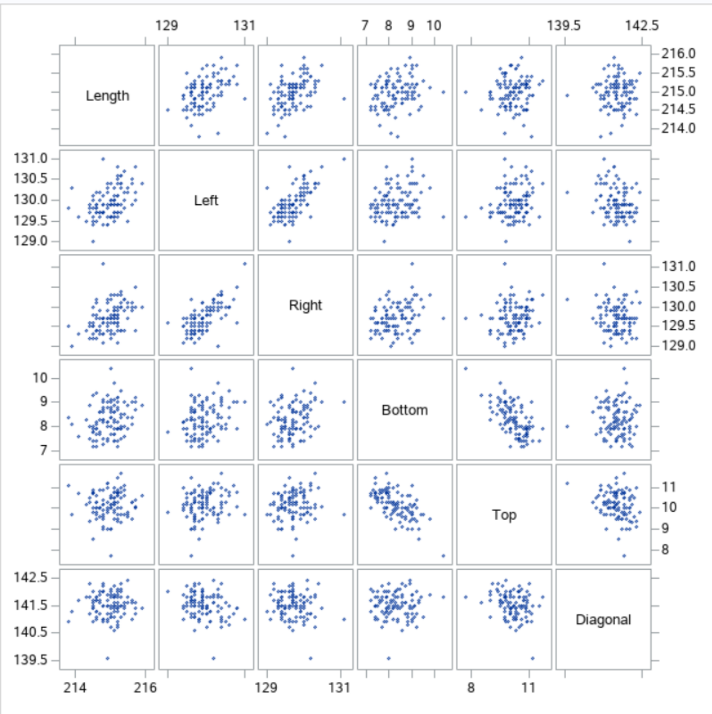
Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Length	100	214.96900	0.38761	21497	213.80000	215.90000
Left	100	129.94300	0.36411	12994	129.00000	131.00000
Right	100	129.72000	0.35533	12972	129.00000	131.10000
Bottom	100	8.30500	0.64281	830.50000	7.20000	10.40000
Top	100	10.16800	0.64899	1017	7.70000	11.70000
Diagonal	100	141.51700	0.44700	14152	139.60000	142.40000

Pearson Correlation Coefficients, N = 100
Prob > |r| under H0: Rho=0

	Length	Left	Right	Bottom	Top	Diagonal
Length	1.00000	0.41105 <.0001	0.41598 <.0001	0.22928 0.0218	0.05745 0.5702	0.03164 0.7547
Left	0.41105 <.0001	1.00000	0.66392 <.0001	0.24204 0.0153	0.20764 0.0382	-0.26458 0.0078
Right	0.41598 <.0001	0.66392 <.0001	1.00000	0.25472 0.0105	0.13289 0.1875	-0.14970 0.1371
Bottom	0.22928 0.0218	0.24204 0.0153	0.25472 0.0105	1.00000	-0.63156 <.0001	-0.00065 0.9949
Top	0.05745 0.5702	0.20764 0.0382	0.13289 0.1875	-0.63156 <.0001	1.00000	-0.25960 0.0091
Diagonal	0.03164 0.7547	-0.26458 0.0078	-0.14970 0.1371	-0.00065 0.9949	-0.25960 0.0091	1.00000

Scatter plot



4) Produce the correlation matrix and an associated Scatterplot of the inputted data for the **forged /fake notes**.

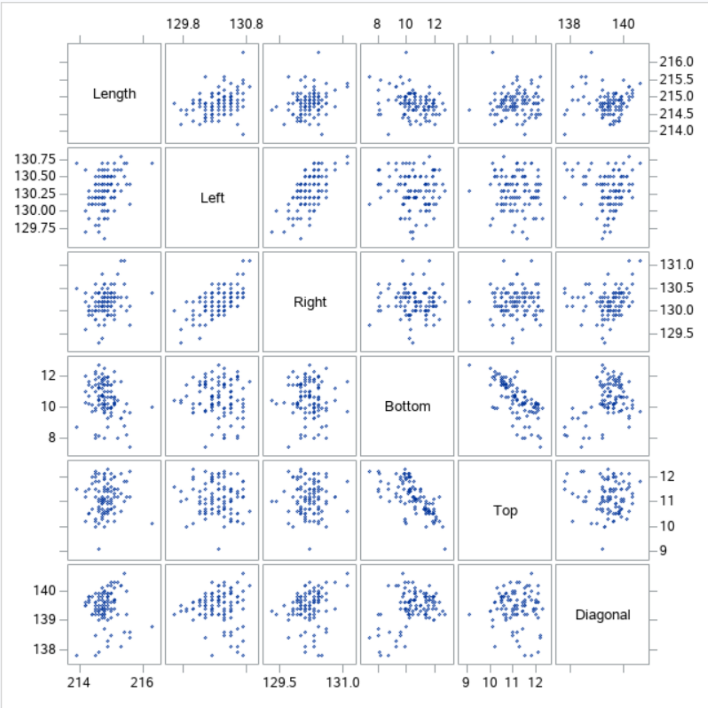
The CORR Procedure

6 Variables:	Length Left Right Bottom Top Diagonal

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Length	100	214.82300	0.35215	21482	213.90000	216.30000
Left	100	130.30000	0.25505	13030	129.60000	130.80000
Right	100	130.19300	0.29823	13019	129.30000	131.10000
Bottom	100	10.53000	1.13195	1053	7.40000	12.70000
Top	100	11.13300	0.63597	1113	9.10000	12.30000
Diagonal	100	139.45000	0.55786	13945	137.80000	140.60000

Pearson Correlation Coefficients, N = 100 Prob > |r| under H0: Rho=0						
	Length	Left	Right	Bottom	Top	Diagonal
Length	1.00000	0.35088 0.0003	0.22853 0.0222	-0.25236 0.0113	0.08678 0.3906	0.05887 0.5607
Left	0.35088 0.0003	1.00000	0.61485 <.0001	-0.08327 0.4101	-0.07348 0.4675	-0.03550 0.7259
Right	0.22853 0.0222	0.61485 <.0001	1.00000	-0.05503 0.5866	0.00070 0.9945	0.20552 0.0402
Bottom	-0.25236 0.0113	-0.08327 0.4101	-0.05503 0.5866	1.00000	-0.68093 <.0001	0.37766 0.0001
Top	0.08678 0.3906	-0.07348 0.4675	0.00070 0.9945	-0.68093 <.0001	1.00000	-0.06221 0.5386
Diagonal	0.05887 0.5607	-0.03550 0.7259	0.20552 0.0402	0.37766 0.0001	-0.06221 0.5386	1.00000

Scatter Plot



5) Run the discriminant analysis using the SAS which allocates a bank note with the following characteristics X_0^T = (214.9, 130.1, 129.9, 9, 10.6, 140.5) to the appropriate grouping i.e., allocates it to either the **genuine** or the **forged/fake class.**

The DISCRIM Procedure

Total Sample Size	200	DF Total	199
Variables	6	DF Within Classes	198
Classes	2	DF Between Classes	1

Number of Observations Read	200
Number of Observations Used	200

Class Level Information					
Status	Variable Name	Frequency	Weight	Proportion	Prior Probability
counterfeit	counterfeit	100	100.0000	0.500000	0.000000
genuine	genuine	100	100.0000	0.500000	1.000000

Within Covariance Matrix Information		
Status	Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
counterfeit	6	-10.79076
genuine	6	-11.21447
Pooled	6	-10.36654

The DISCRIM Procedure

Generalized Squared Distance to Status		
From Status	counterfeit	genuine
counterfeit	.	43.64525
genuine	.	-11.21447

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.BANK
Resubstitution Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into Status			
From Status	counterfeit	genuine	Total
counterfeit	0 0.00	100 100.00	100 100.00
genuine	0 0.00	100 100.00	100 100.00
Total	0 0.00	200 100.00	200 100.00
Priors	0	1	

Error Count Estimates for Status			
	counterfeit	genuine	Total
Rate	1.0000	0.0000	0.0000
Priors	0.0000	1.0000	

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
121.899123	21	<.0001

Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.
Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.BANK
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into Status			
From Status	counterfeit	genuine	Total
counterfeit	0 0.00	100 100.00	100 100.00
genuine	0 0.00	100 100.00	100 100.00
Total	0 0.00	200 100.00	200 100.00
Priors	0	1	

Error Count Estimates for Status			
	counterfeit	genuine	Total
Rate	1.0000	0.0000	0.0000
Priors	0.0000	1.0000	

The DISCRIM Procedure
Classification Summary for Test Data: WORK.TEST
Classification Summary using Quadratic Discriminant Function

Observation Profile for Test Data	
Number of Observations Read	1
Number of Observations Used	1

Number of Observations and Percent Classified into Status			
	counterfeit	genuine	Total
Total	0 0.00	1 100.00	1 100.00
Priors	0	1	

6) Using the SAS DISCRIM and resultant output answer the following questions.

a) Is $\sum_1 = \sum_2$?

Ans:

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
121.899123	21	<.0001

Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.
Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

We know that this as the chi-squared test seen above returned a p-value of < 0.0001, meaning we should reject the null hypothesis (that the two covariance matrices are equal), and instead accept that the two covariance matrices are not equal.

b) How is the bank note with X_0^T = (214.9, 130.1, 129.9, 9, 10.6, 140.5) allocated?

Ans:

Obs	length	left	right	bottom	top	diag	fake	real	_INTO_
1	214.9	130.1	129.9	9	10.6	140.5	.000002526	1.00000	real

Number of Observations and Percent Classified into Status			
	counterfeit	genuine	Total
Total	0 0.00	1 100.00	1 100.00
Priors	0	1	

Appendix:

```
/*1*/

data thc;

	infile '/home/s37007540/ma/ma ass 2/THC.csv' delimiter="," firstobs=2;

	input varieties chem1 chem2 chem3 chem4 chem5 chem6 chem7 chem8 chem9 chem10 chem11 chem12 chem13;

run;


/*1.1*/

proc means data=thc ;

var varieties chem1 chem2 chem3 chem4 chem5 chem6 chem7 chem8 chem9 chem10 chem11 chem12 chem13;

run;


proc means data=thc std ;

var varieties chem1 chem2 chem3 chem4 chem5 chem6 chem7 chem8 chem9 chem10 chem11 chem12 chem13;

run;


/*1.2*/

proc princomp data=thc ;
```

```
run;

proc sgscatter data=thc;
matrix chem1 chem2 chem3 chem4 chem5 chem6 chem7 chem8 chem9 chem10 chem11 chem12 chem13;
run;

/*1.3.a*/
proc princomp data=thc cov;
run;

/*1.4*/
proc princomp data=thc;
var chem1 chem2 chem3 chem4 chem5 chem6 chem7 chem8 chem9 chem10 chem11 chem12 chem13;
run;



/*2*/

DATA SocioEconomics;
input Population School Employment Services HouseValue;
datalines;
5700 12.8 2500 270 25000
1000 10.9 600 10 10000
3400 8.8 1000 10 9000
3800 13.6 1700 140 25000
4000 12.8 1600 140 25000
8200 8.3 2600 60 12000
1200 11.4 400 10 16000
9100 11.5 3300 60 14000
9900 12.5 3400 180 18000
```

```
9600 13.7 3600 390 25000
9600 9.6 3300 80 12000
9400 11.4 4000 100 13000
;

proc means data=SocioEconomics mean std maxdec=4;
var Population School Employment Services HouseValue ;
run;



proc factor data=SocioEconomics simple corr;
run;

proc factor data=SocioEconomics n=3 simple corr;
run;
/*2.5*/
proc princomp data=SocioEconomics;
run;



/*2.6*/
proc factor data=SocioEconomics n=5 score;
run;





/*3*/
data bank;
informat id 2.0 Status $20.0 Length 2.0 Left 2.0 Right 2.0 Bottom 2.0 Top 2.0 Diagonal 2.0;
infile "/home/s37007540/ma/ma ass 2/banknote.csv" delimiter="," firstobs=2 dsd missover;
```

```
input  id status $ length left right bottom top diagonal;
run;


proc sql;
create table genuine as
select*from bank
where status eq "genuine";
quit;


/*mean of genuine*/*/
proc means data=genuine;
var length left right bottom top diagonal;
run;

 /*mean of genuine*/
proc means data=genuine mean;
var length left right bottom top diagonal;
run;

proc sql;
create table fake as
select*from bank
where status eq "counterfeit";
quit;

/*mean of fake*/*/
proc means data=fake;
var length left right bottom top diagonal;
run;
```

```
/*mean of fake*/
proc means data=fake ;
var length left right bottom top diagonal;
run;
```

304

```
/*variance- covariance matrix of genuine notes*/
proc corr data = genuine cov;
var length left right bottom top diagonal;
run;

/*variance- covariance matrix of fake notes*/
proc corr data = fake cov;
var length left right bottom top diagonal;
run;

/*correlation matrix of genuine notes*/
proc corr data=genuine;
var length left right bottom top diagonal;
run;

/*scatter plot matrix of genuine notes*/
proc sgscatter data=genuine;
matrix length left right bottom top diagonal;
run;

/*correlation matrix  of fake notes*/

proc corr data=fake;
var length left right bottom top diagonal;
```

```
run;

/*scatter plot matrix of fake notes*/
proc sgscatter data=fake;
matrix length left right bottom top diagonal;
run;

/*uploading test data*/
data test;
input length left right bottom top diagonal;
cards;
214.9 130.1 129.9 9 10.6 140.5
;
run;

/*discriminant analysis on test data*/
proc discrim data=bank
 pool=test
 crossvalidate
 testdata=test
 testout=a;
 class status;
 var length left right bottom top diagonal;
 prior "genuine"=0.99 "fake"=0.01;
```

fa data

1. run; Prepare the dataset for a Factor analysis via SAS.

Ans:

```
data SocioEconomics;
input Population School Employment Services HouseValue;
datalines;
5700 12.8 2500 270 25000
1000 10.9 600 10 10000
3400 8.8 1000 10 9000
3800 13.6 1700 140 25000
4000 12.8 1600 140 25000
8200 8.3 2600 60 12000
1200 11.4 400 10 1600
9100 11.5 3300 60 14000
9900 12.5 3400 180 18000
9600 13.7 3600 390 25000
9600 9.6 3300 80 12000
9400    4 4000 100 13000
;
```