STAT448 - Advanced Data Analysis

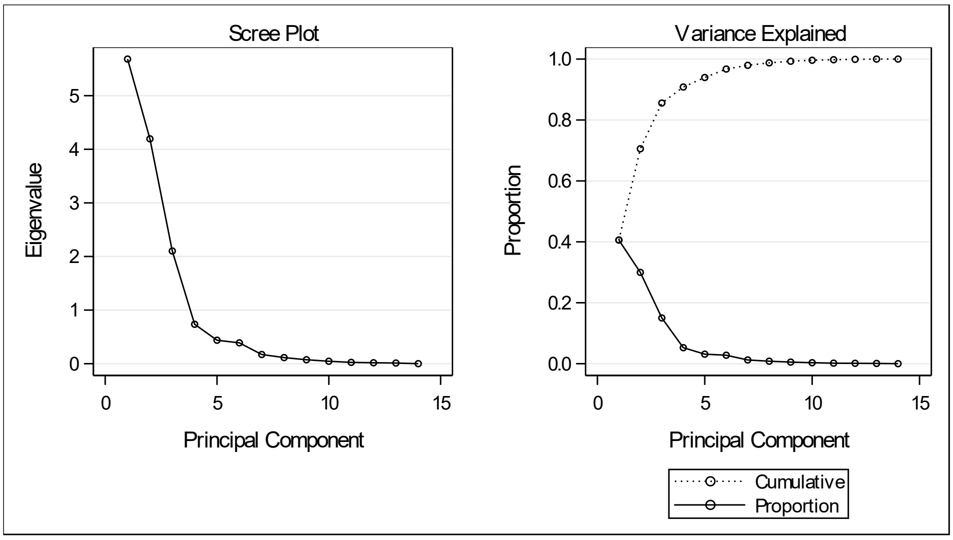
Homework 5

Name: Xinyan Yang

Exercise 1 Solution:

(a).

Eigenvalues of the Correlation Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	5.68286683	1.48810625	0.4059	0.4059
2	4.19476058	2.09269357	0.2996	0.7055
3	2.10206700	1.36652186	0.1501	0.8557
4	0.73554515	0.29785430	0.0525	0.9082
5	0.43769085	0.04886854	0.0313	0.9395
6	0.38882230	0.21758944	0.0278	0.9673
7	0.17123286	0.05740175	0.0122	0.9795
8	0.11383110	0.04044157	0.0081	0.9876
9	0.07338953	0.02805442	0.0052	0.9929
10	0.04533511	0.02073165	0.0032	0.9961
11	0.02460346	0.00710262	0.0018	0.9979
12	0.01750084	0.00538929	0.0013	0.9991
13	0.01211155	0.01186871	0.0009	1.0000
14	0.00024283		0.0000	1.0000



From the above results, I would keep **3** components to retain at least 85% of the total variation from the original variables. If based on the average eigenvalue, I would keep **3** components too. If based on the scree plot, I would also choose to keep 3 components.

(b).

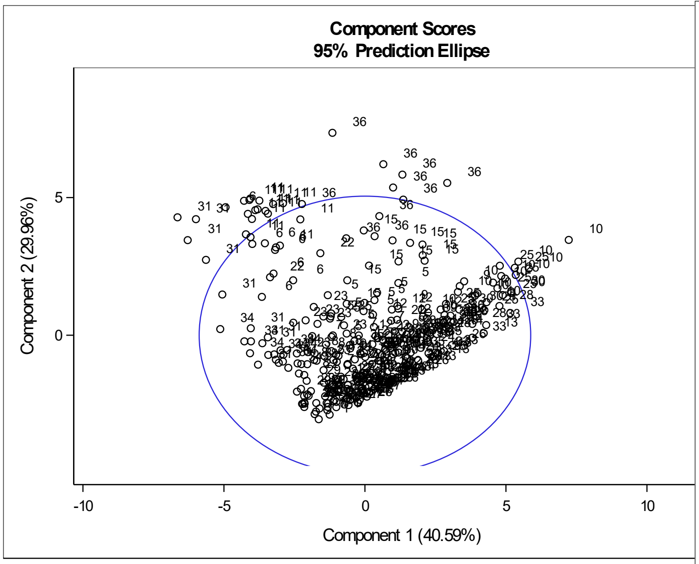
Eigenvectors								
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8
Eccentricity	-.093812	-.192439	0.538282	0.129275	-.172358	0.617888	0.009354	0.003232
AspectRatio	-.190165	-.025256	0.519236	-.168502	0.463153	-.470015	0.177409	0.371746
Elongation	-.226596	0.179998	0.488936	-.023459	-.304155	0.084670	0.072061	-.075301
Solidity	0.185012	-.408405	0.135090	0.012671	0.112225	0.034265	0.245483	0.058684
StochasticConvexity	0.159994	-.382523	0.169293	-.026241	0.479106	0.060754	-.575300	-.394742
IsoperimetricFactor	0.206267	-.348799	-.251244	0.116345	0.173495	0.333501	0.313804	0.411369
MaximalIndentationDepth	-.194033	0.403688	-.076036	0.062834	0.322389	0.216907	-.084478	-.162308
Lobedness	-.214965	0.356621	-.089326	0.028774	0.486409	0.416836	0.038788	0.214039
AverageIntensity	0.372282	0.200126	0.119865	-.095298	0.014949	0.036719	-.009030	0.081528
AverageContrast	0.365666	0.197444	0.130476	0.178703	-.023820	-.023470	-.054061	0.009304
Smoothness	0.360162	0.203690	0.139989	0.228768	0.070729	-.040470	0.124641	-.055633
ThirdMoment	0.317546	0.188619	0.138098	0.538522	0.114144	-.120489	0.163919	-.141813
Uniformity	0.305584	0.124336	0.037886	-.679273	0.098205	0.162233	0.410430	-.382284
Entropy	0.348165	0.182884	0.079373	-.300691	-.148457	0.094169	-.501037	0.528581

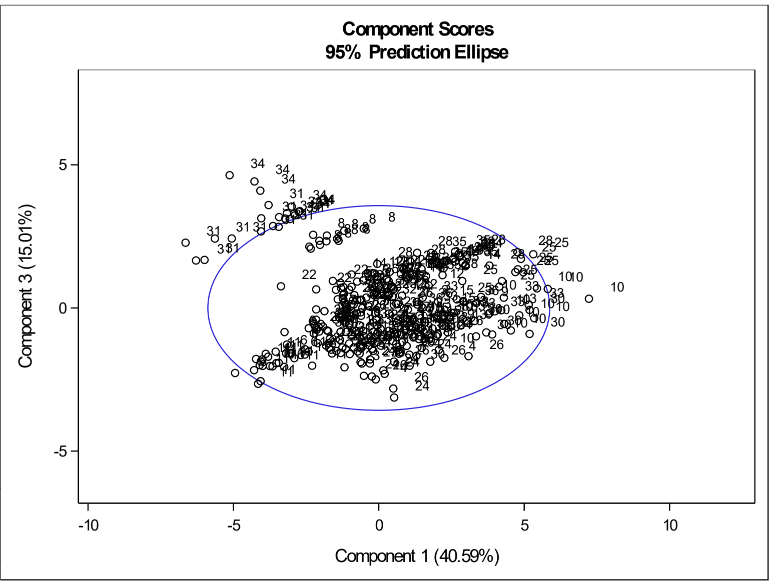
We can get the following conclusions from the above table. The first component has large positive coefficient on the last six features and these are texture features measuring the intensity, therefore **the first PC picks up on texture features like intensity.**

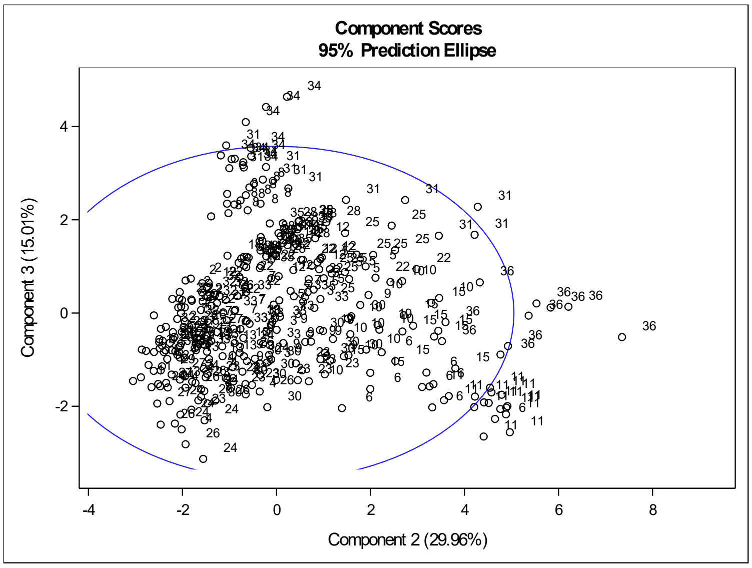
The second component has large negative coefficients on solidity, stochasticconvexity and isoperimetricfactor which measure the convexity of leaves, and it also has large positive coefficients on maximalIndentationDepth and Lobedness which measure the indentation. **Therefore the second PC picks up on shape features like convexity and indentation, and large positive value means more indentation and less convexity.**

The third component has large positive coefficients on eccentricity, aspectratio and elongation which measure some characteristics of ellipse, **therefore the third PC picks up on shape features like slenderness**.

(c)







From the scree plots we can get the following conclusions.

Species 10 has extreme large value and species 31 has extreme small value on component 1, therefore species 10 has large intensity and species 31 has small intensity.

Species 36 and 11 have extreme large value on component 2, they have many indentations and less convexity with their shape;

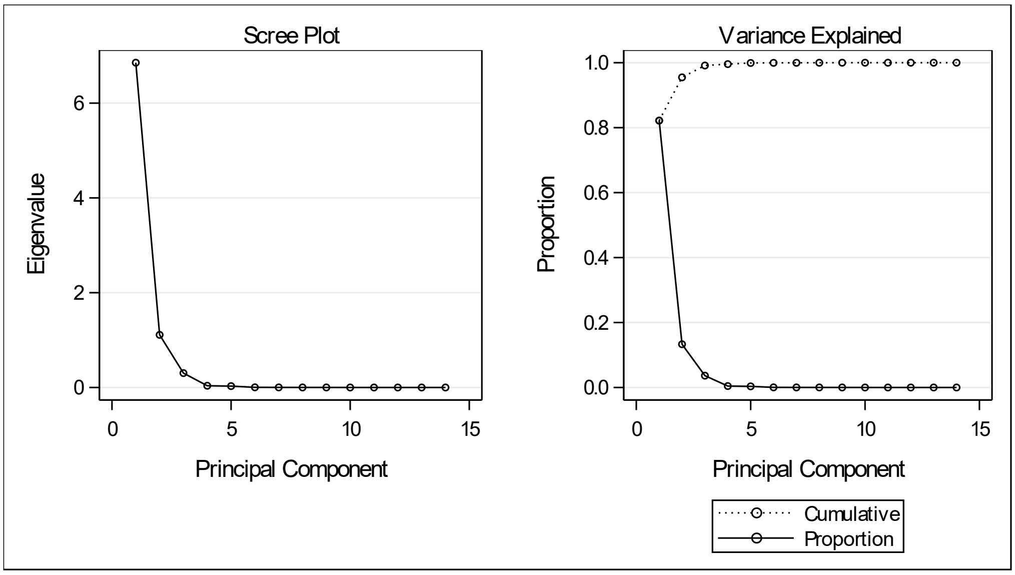
Species 34, 31 and 8 have extreme large value on component 3, they tend to have slender shape.

Exercise 2 Solution:

(a).

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	6.85274257	5.74250615	0.8218	0.8218
2	1.11023642	0.80664635	0.1331	0.9549
3	0.30359007	0.26713214	0.0364	0.9913
4	0.03645793	0.00745915	0.0044	0.9957
5	0.02899878	0.02578640	0.0035	0.9992
6	0.00321239	0.00151672	0.0004	0.9996
7	0.00169566	0.00060366	0.0002	0.9998
8	0.00109200	0.00026536	0.0001	0.9999
9	0.00082665	0.00076185	0.0001	1.0000
10	0.00006480	0.00002701	0.0000	1.0000
11	0.00003779	0.00003379	0.0000	1.0000
12	0.00000400	0.00000392	0.0000	1.0000

Eigenvalues of the Covariance Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
13	0.00000008	0.00000007	0.0000	1.0000
14	0.00000001		0.0000	1.0000



Based on the total variation that PC can explain, I would keep 2 components to retain at least 85% variation. Based on the eigenvalues, I would also choose 2 components. Based on scree plot, 3 PCs would be kept.

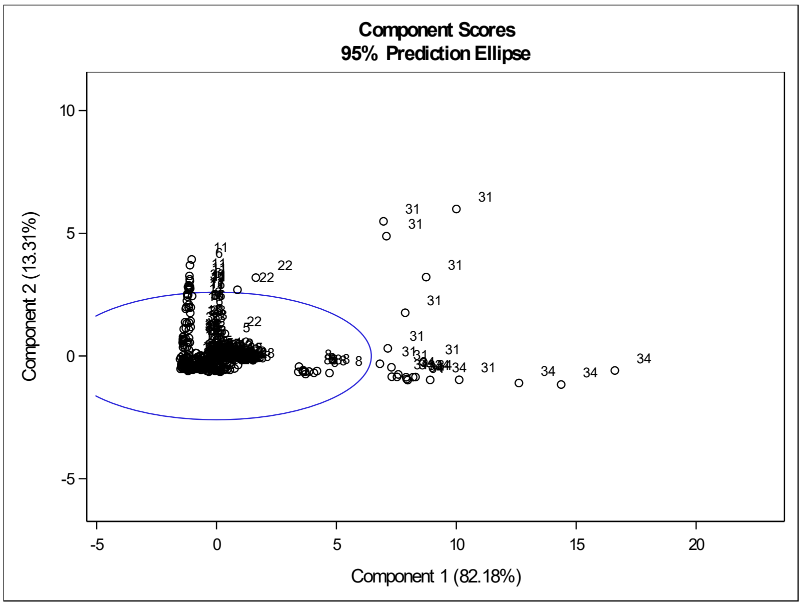
(b).

Eigenvectors								
	Prin1	**Prin2**	**Prin3**	**Prin4**	**Prin5**	**Prin6**	**Prin7**	**Prin8**
Eccentricity	0.043578	-.059491	-.049223	0.540684	0.712685	0.089212	-.410460	0.123592
AspectRatio	0.992322	-.070973	0.065926	-.072444	-.012560	-.017670	-.013510	-.000246
Elongation	0.051542	0.058718	0.030490	0.638389	-.025286	-.293785	0.632365	-.307770
Solidity	-.000792	-.091540	-.024576	-.121592	0.261341	0.057878	0.534778	0.686743
StochasticConvexity	0.003775	-.080453	-.013748	-.174329	0.296226	0.759101	0.337885	-.409900
IsoperimetricFactor	-.040427	-.117221	-.087745	-.488549	0.546607	-.567745	0.115311	-.288148
MaximalIndentationDepth	0.001738	0.034492	0.007448	0.018635	-.035985	0.048732	-.029575	0.062847
Lobedness	0.059659	0.973906	0.085612	-.098892	0.168535	0.017384	0.027714	0.020145
AverageIntensity	-.004115	-.005720	0.056866	-.007066	0.008189	-.007060	0.062783	0.167464
AverageContrast	-.006208	-.008331	0.073973	0.005471	0.010883	0.009626	0.103785	0.347020
Smoothness	-.001462	-.001894	0.018970	-.000988	0.003674	-.002269	0.030082	0.103210
ThirdMoment	-.000496	-.000561	0.005667	0.000179	0.001385	0.000382	0.010510	0.048556
Uniformity	-.000040	-.000070	0.000578	-.000359	0.000056	-.000181	0.000726	0.000383
Entropy	-.073976	-.097949	0.983527	-.028453	0.080961	-.025994	-.025133	-.038816

The first component has large positive coefficient on aspectratio, which measures the shape of a leaf.

The second component has large positive coefficient on lobedness, which measures how lobed a leaf is.

(c).



Species 34, 31 have large positive value on component 1, which means that these species have a really elongated shape;

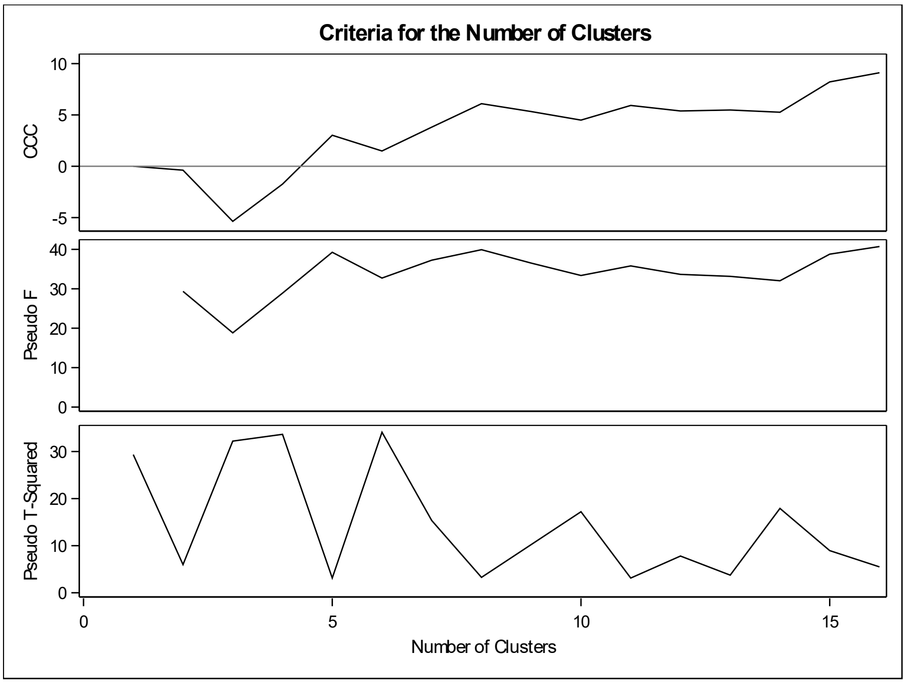
Species 11 and 6 have large positive value on component2, which means that these species have more identations.

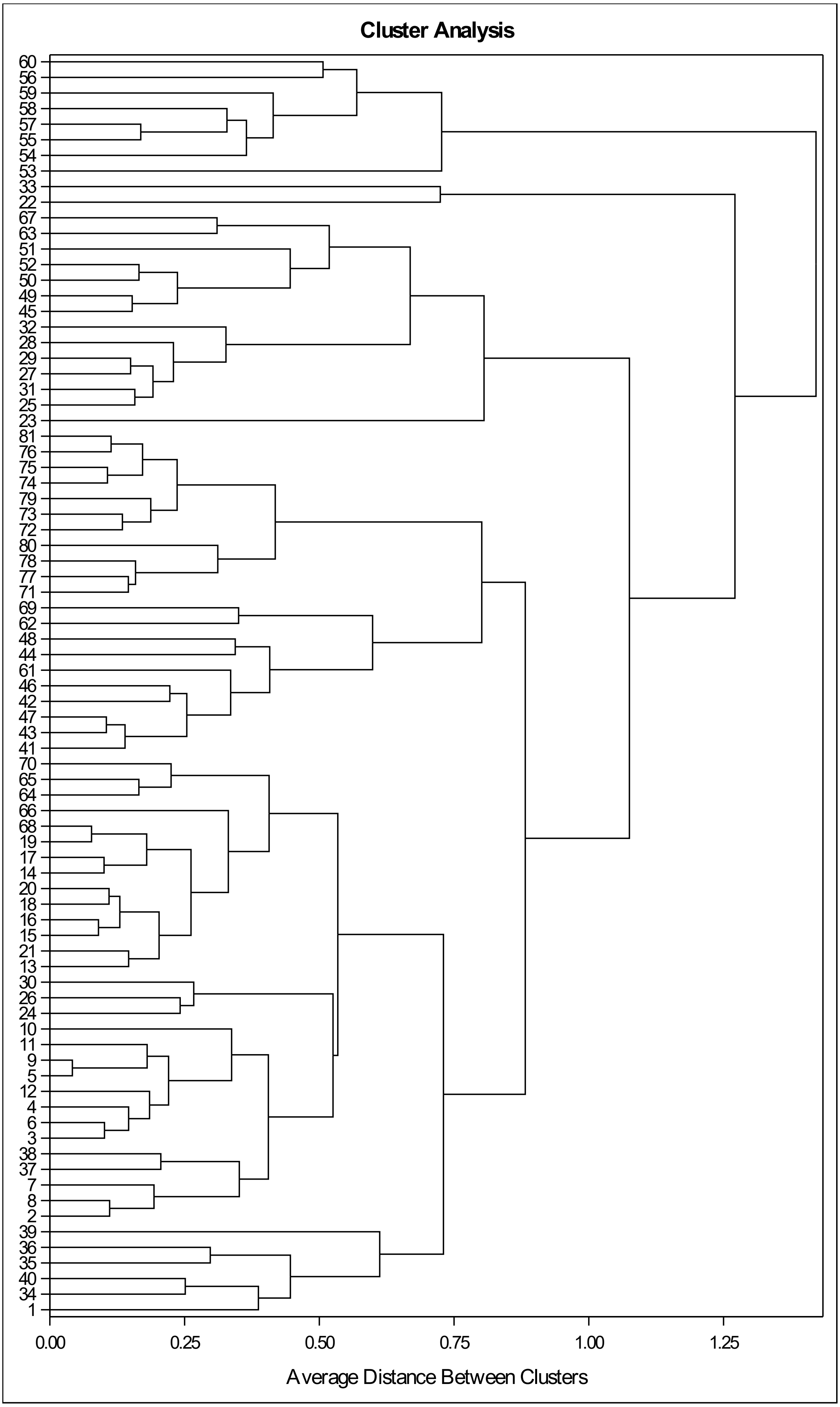
(d).

The covariance-based PCA result give extremely large coefficient to one predictor in a component because of that particular predictor's large covariance. The correlation-based PCA result has kind of balanced coefficient among many predictors and won't give great weight to a single one predictor, which is more reasonable.

Exercise 3 Solution:

(a).





Cluster Analysis
60
56
59
58
57
55
54
53
33
22
67
63
51
52
50
49
45
32
28
29
27
31
25
23
81
76
75
74
79
73
72
80
78
77
71
69
62
48
44
61
46
42
47
43
41
70
65
64
66
68
19
17
14
20
18
16
15
21
13
30
26
24
10
11
9
5
12
4
6
3
38
37
7
8
2
39
36
35
40
34
1
0.00
0.25
0.50
0.75
1.00
1.25
Average Distance Between Clusters

Based on the ccc statistics, we should choose 8, 11 or 16 clusters. Based on the pseudo F statistic, we should choose 5, 8 or 16 clusters. Based on pseudo t^2 statistic, we should choose 2, 5, 8, 11, 13 or 16 clusters. Based on the dendrogram, we should choose a cut-off value around 0.75 at which clusters have a comparatively large distance.

Therefore I decide to choose 8 clusters.

(b).

Table of CLUSTER by Species									
CLUSTER	**Species**								
Frequency	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Total**
1	11	9	3	2	0	0	5	0	30
2	0	0	0	0	7	0	3	0	10
3	0	0	0	0	0	0	0	11	11
4	0	0	6	0	5	0	2	0	13
5	0	0	0	0	0	8	0	0	8
6	1	0	0	5	0	0	0	0	6
7	0	1	0	1	0	0	0	0	2
8	0	0	1	0	0	0	0	0	1
Total	12	10	10	8	12	8	10	11	81

From the above table we can see that, species 6 and 8 are separated out pretty well. Species 4 is OK. And species 1, 2, 3 and 7 are grouped together in cluster1. Species 3 and 5 are grouped together in cluster4. Species 5 and 7 are grouped together in cluster2.

We can find that species 6 and 8 are very leaves with very special characteristics, species6 is lobed and species 8 is elongated, which make these two easily to stand out.

Species1, 2, 3 and 7 which in cluster1 show a ellipse shape, which make them hard to identify.

Species 3 and 5 in cluster4 show a triangle shape both.

Species 5 and 7 in cluster2 show a rough edge both.