OPIM 5604 B15 – Predictive Modeling Assignment 3 Meghana Kasula (Net ID=mek15120)

"The work contained and presented here is my work and my work alone."

4.1 Breakfast Cereals. Use the data for the breakfast cereals example in Section 4.8 (Cereals.jmp) to explore and summarize the data as follows:

a. Which variables are continuous/numerical? Which are ordinal? Which are nominal?

Continuous variables-calories protein, fat, sodium, fiber, carbo, sugars, potass, vitamins, shelf, weight, cups, rating.

Nominal variables - name, mfr and type.

b. Calculate the following summary statistics: mean, median, min, max, and standard deviation for each of the continuous variables, and the count for each categorical variable. This can be done using *Cols* > *Columns Viewer*.

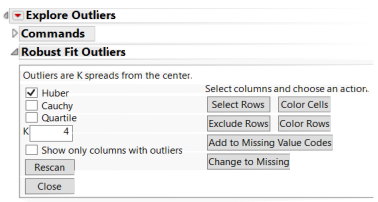
13 Columns Clear Select Distribution

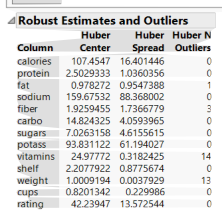
Columns	N	N Missing	Min	Max	Mean	Std Dev	Median	Lower Quartile	Upper Quartile	Interquartile Range
calories	77	0	50	160	106.88311688	19.484119057	110	100	110	10
protein	77	0	1	6	2.5454545455	1.0947897484	3	2	3	1
fat	77	0	0	5	1.012987013	1.0064725595	1	0	2	2
sodium	77	0	0	320	159.67532468	83.83229524	180	127.5	215	87.5
fiber	77	0	0	14	2.1519480519	2.3833639644	2	0.5	3	2.5
carbo	76	1	5	23	14.802631579	3.9073255537	14.5	12	17	5
sugars	76	1	0	15	7.0263157895	4.3786563668	7	3	11	8
potass	75	2	15	330	98.666666667	70.41063597	90	40	120	80
vitamins	77	0	0	100	28.246753247	22.342522501	25	25	25	0
shelf	77	0	1	3	2.2077922078	0.8325241001	2	1	3	2
weight	77	0	0.5	1.5	1.0296103896	0.1504767997	1	1	1	0
cups	77	0	0.25	1.5	0.821038961	0.2327161384	0.75	0.67	1	0.33
rating	77	0	18.042851	93.704912	42.665704987	14.047288744	40.400208	32.690838	51.2102925	18.5194545

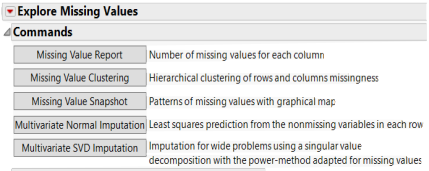
i. Is there any evidence of extreme values?

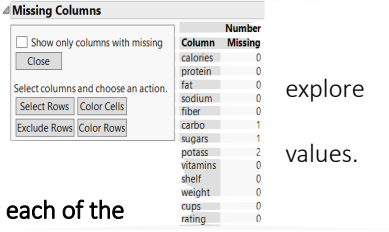
Yes, if we explore outliers through col>model utilities > explore outliers.

We can see that vitamins, fat and weight has outliers which has been calculated using Huber spread and from Huber center.









ii. Which, if any, of the variables is missing values?

If we explore outliers through col>model utilities > explore missing values.

We can see that potass , sugar and carbo has missing values.

c. Use *Analyze* > *Distribution* to plot a histogram for each of the continuous variables and create summary statistics. Based on the histograms and summary statistics, answer the following questions:

i. Which variables have the largest variability?

We can calculate the variability using coefficient of variance. Here coefficient of variance is particularly important since we are comparing data of two different sets, like potass and fat etc. this can give us the clear idea about variability in all measures. It's formula is-

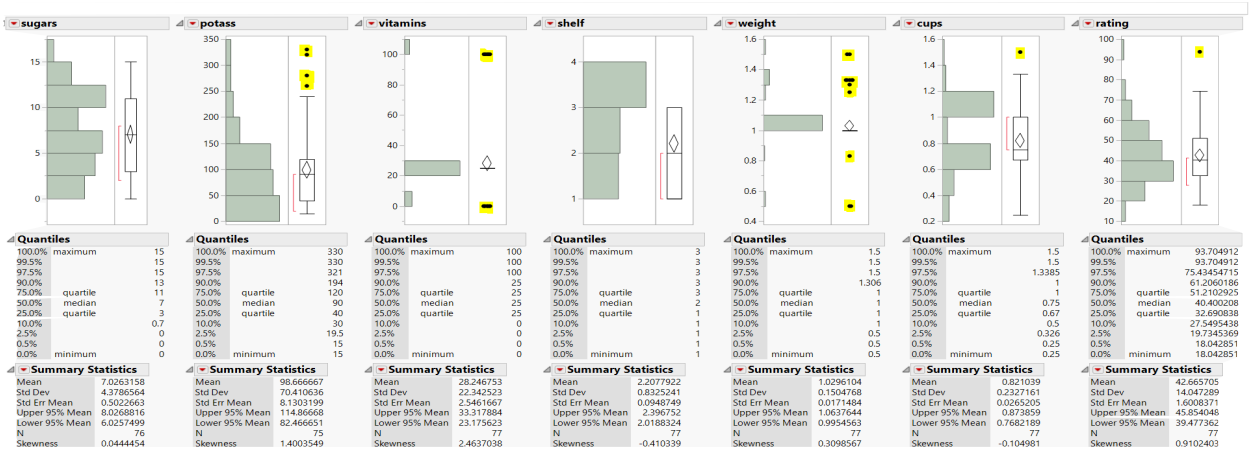
Coefficient of variance = standard Deviation / Mean

Hence, after conducting the formula for each of the measures, I get to know that the vitamins have the highest coefficient of variance. Which is 0.788.

ii. Which variables seem skewed?

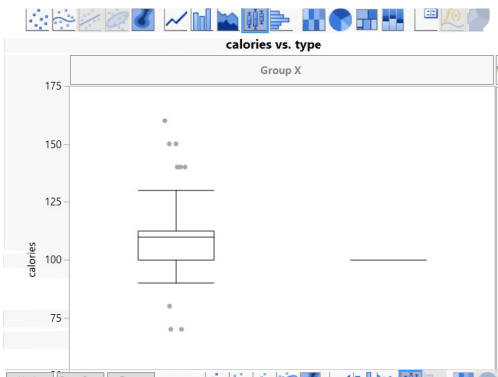
iii. Are there any values that seem extreme?

Rule of thumb is that acceptable amount of



skewness is between -1 to +1. Hence I have highlighted the values which are either lesser than -1 and more than 1. So, very skewed values are fat, fiber, potass and vitamins.

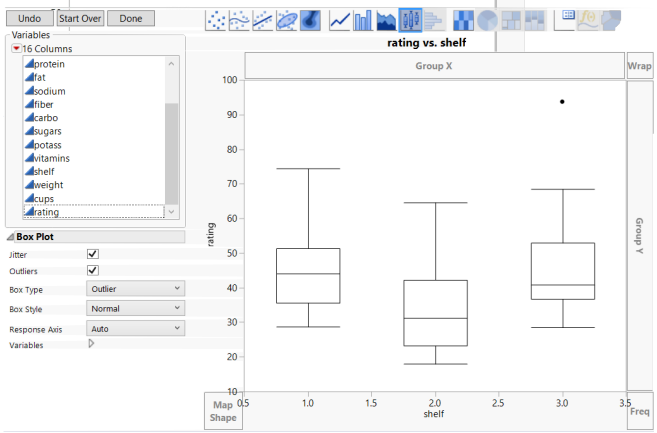
Yes, there are extreme values! I have highlighted all the extreme values in Yellow. Those are the outliers.



d. Use the *Graph Builder* to plot a side-by-side boxplot comparing the calories in hot vs. cold cereals. What does this plot show us?

The plot shows us that there is a large range in Cold cereals ranging from 50 -160. It also includes outliers.

Hot cereals have only one value of calories, which is 100. So the mean, median and mode has the same value.



e. Use the *Graph Builder* to plot a side-by-side boxplot of consumer rating as a

function of the shelf height (the variable *shelf*). If we were to predict consumer rating from shelf height, does it appear that we need to keep all three categories of shelf height?

As we look at the box plot on the right, we can observe that the plot at shelf 1 and shelf 3 are similar looking. Moreover, the plot at shelf 3 looks like a subset of plot 1. Hence, we can conclude combining the shelf 3 and 1 would not make much difference. We can effectively reduce the numbers of shelves to two.

f. Compute the correlation table and generate a scatterplot matrix for the continuous variables (use *Analyze* > *Multivariate Methods* > *Multivariate*).

i. Which pair of variables is most strongly correlated?

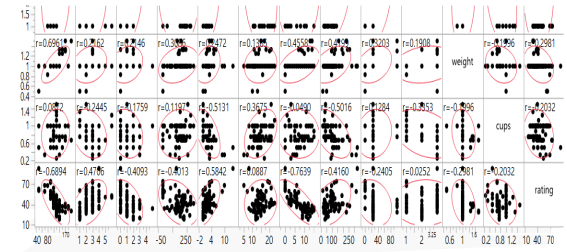
Correlations

	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
calories	1.0000	0.0191	0.4986	0.3006	-0.2934	0.2576	0.5665	-0.0721	0.2654	0.0972	0.6961	0.0872	-0.6894
protein	0.0191	1.0000	0.2084	-0.0547	0.5003	-0.0250	-0.2919	0.5637	0.0073	0.1339	0.2162	-0.2445	0.4706
fat	0.4986	0.2084	1.0000	-0.0054	0.0167	-0.3000	0.3025	0.2004	-0.0312	0.2637	0.2146	-0.1759	-0.4093
sodium	0.3006	-0.0547	-0.0054	1.0000	-0.0707	0.2977	0.0589	-0.0426	0.3615	-0.0697	0.3086	0.1197	-0.4013
fiber	-0.2934	0.5003	0.0167	-0.0707	1.0000	-0.3804	-0.1388	0.9115	-0.0322	0.2975	0.2472	-0.5131	0.5842
carbo	0.2576	-0.0250	-0.3000	0.2977	-0.3804	1.0000	-0.4712	-0.3650	0.2192	-0.1926	0.1385	0.3675	0.0887
sugars	0.5665	-0.2919	0.3025	0.0589	-0.1388	-0.4712	1.0000	0.0014	0.0982	0.0684	0.4558	-0.0490	-0.7639
potass	-0.0721	0.5637	0.2004	-0.0426	0.9115	-0.3650	0.0014	1.0000	-0.0054	0.3858	0.4199	-0.5016	0.4160
vitamins	0.2654	0.0073	-0.0312	0.3615	-0.0322	0.2192	0.0982	-0.0054	1.0000	0.2993	0.3203	0.1284	-0.2405
shelf	0.0972	0.1339	0.2637	-0.0697	0.2975	-0.1926	0.0684	0.3858	0.2993	1.0000	0.1908	-0.3353	0.0252
weight	0.6961	0.2162	0.2146	0.3086	0.2472	0.1385	0.4558	0.4199	0.3203	0.1908	1.0000	-0.1996	-0.2981
cups	0.0872	-0.2445	-0.1759	0.1197	-0.5131	0.3675	-0.0490	-0.5016	0.1284	-0.3353	-0.1996	1.0000	-0.2032
rating	-0.6894	0.4706	-0.4093	-0.4013	0.5842	0.0887	-0.7639	0.4160	-0.2405	0.0252	-0.2981	-0.2032	1.0000

There are 2 missing values. The correlations are estimated by Pairwise method

As we analyze the correlations matrix above, the highest correlation seems to be between the pair fiber and potass. They are highlighted in blue.

ii. How can we reduce the number of variables based on these correlations?



Principal Components / Factor Analysis

Principal Components: on Correlations

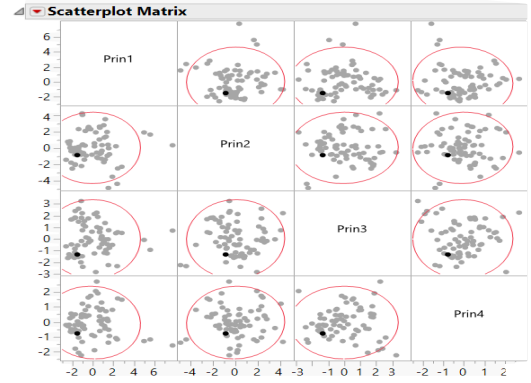
Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	3.6043	27.725		27.725
2	3.1661	24.355		52.080
3	1.8723	14.402		66.482
4	1.0911	8.393		74.875
5	0.9682	7.448		82.322
6	0.7210	5.546		87.868
7	0.6623	5.094		92.963
8	0.4373	3.364		96.326
9	0.2992	2.301		98.628
10	0.0917	0.705		99.333
11	0.0581	0.447		99.780
12	0.0241	0.186		99.966
13	0.0045	0.034		100.000

Multivariate

Correlations

	Prin1	Prin2	Prin3	Prin4
Prin1	1.0000	0.0103	-0.0007	-0.0463
Prin2	0.0103	1.0000	0.0202	0.0331
Prin3	-0.0007	0.0202	1.0000	0.0049
Prin4	-0.0463	0.0331	0.0049	1.0000

Scatterplot Matrix



As we select on the red triangle beside principal components and save principal components “4”, we get 4 new columns based on the correlations where all the measures are reduced.

The second picture represents the correlation matrix of those reduced variables. This is how we can reduce number of variables using correlations.

iii. How would the correlations change if we normalized the data first?

Correlations

	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
calories	1.0000	0.0191	0.4986	0.3006	-0.2934	0.2576	0.5665	-0.0721	0.2654	0.0972	0.6961	0.0872	-0.6894
protein	0.0191	1.0000	0.2084	-0.0547	0.5003	-0.0250	-0.2919	0.5637	0.0073	0.1339	0.2162	-0.2445	0.4706
fat	0.4986	0.2084	1.0000	-0.0054	0.0167	-0.3000	0.3025	0.2004	-0.0312	0.2637	0.2146	-0.1759	-0.4093
sodium	0.3006	-0.0547	-0.0054	1.0000	-0.0707	0.2977	0.0589	-0.0426	0.3615	-0.0697	0.3086	0.1197	-0.4013
fiber	-0.2934	0.5003	0.0167	-0.0707	1.0000	-0.3804	-0.1388	0.9115	-0.0322	0.2975	0.2472	-0.5131	0.5842
carbo	0.2576	-0.0250	-0.3000	0.2977	-0.3804	1.0000	-0.4712	-0.3650	0.2192	-0.1926	0.1385	0.3675	0.0887
sugars	0.5665	-0.2919	0.3025	0.0589	-0.1388	-0.4712	1.0000	0.0014	0.0982	0.0684	0.4558	-0.0490	-0.7639
potass	-0.0721	0.5637	0.2004	-0.0426	0.9115	-0.3650	0.0014	1.0000	-0.0054	0.3858	0.4199	-0.5016	0.4160
vitamins	0.2654	0.0073	-0.0312	0.3615	-0.0322	0.2192	0.0982	-0.0054	1.0000	0.2993	0.3203	0.1284	-0.2405
shelf	0.0972	0.1339	0.2637	-0.0697	0.2975	-0.1926	0.0684	0.3858	0.2993	1.0000	0.1908	-0.3353	0.0252
weight	0.6961	0.2162	0.2146	0.3086	0.2472	0.1385	0.4558	0.4199	0.3203	0.1908	1.0000	-0.1996	-0.2981
cups	0.0872	-0.2445	-0.1759	0.1197	-0.5131	0.3675	-0.0490	-0.5016	0.1284	-0.3353	-0.1996	1.0000	-0.2032
rating	-0.6894	0.4706	-0.4093	-0.4013	0.5842	0.0887	-0.7639	0.4160	-0.2405	0.0252	-0.2981	-0.2032	1.0000

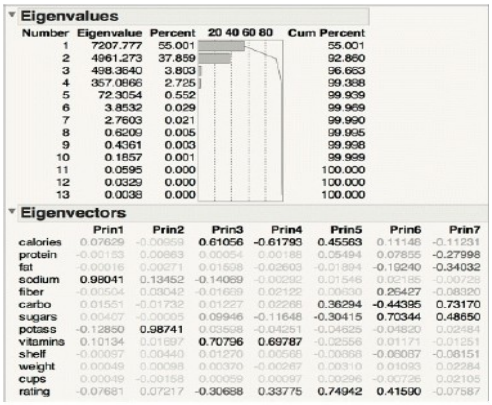
There are 2 missing values. The correlations are estimated by Pairwise method

Correlation matrix is done only after the data is normalized. Hence, there would not be any difference if we differently normalize the data apply the correlation matrix.

The above one is the correlation matrix before standardizing attributes. The beside one is after. We can observe no change at all.

g. Consider the first column on the left under *Eigenvectors* in Figure 4.14. Describe briefly what this column represents.

Here, in first has the



first column, the Eigen vector Prin 1, it is evident that the principal component is dominated by Sodium content. It highest value of 0.98041 at 55.001%.

Figure 4.14 PCA output using all 13 numerical variables in the breakfast cereals dataset. The eigenvectors table gives results for the first seven principal components.

4.2 Chemical Features of Wine. Figure 4.17 shows the PCA output for analyses conducted on normalized (correlations) and non normalized data (covariances).Each variable represent a chemical characteristics of wine, and each case is a different wine.

a. The data are in the file Wine.jmp. Consider the variances in the columns labeled *Eigenvalue* for the principal components analysis conducted using covariance. Why is the variance for the first principal component so much larger than the others?

Here, instead of the correlations we will take covariance. We will have to do the following. Analyze> Multivariate Methods> Principal component > on covariance.

Eigenvectors

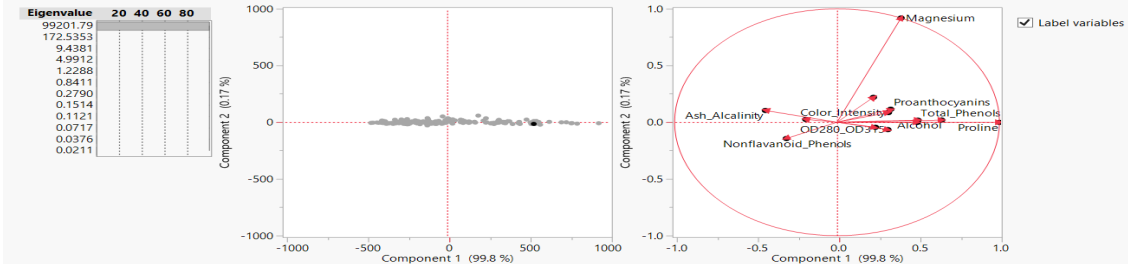
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12	Prin13
Alcohol	0.00166	0.00120	0.01687	0.14145	0.02034	0.19412	0.92328	0.28482	-0.08660	-0.00224	-0.01497	-0.01565	0.00803
Malic_Acid	-0.00068	0.00215	0.12200	0.16039	-0.61288	0.74247	-0.15011	-0.06467	-0.01566	-0.01851	-0.02319	0.06730	-0.01109
Ash	0.00019	0.00459	0.05199	-0.00977	0.02018	0.04175	0.04501	-0.14934	-0.07365	-0.08680	0.95401	-0.13206	-0.17369
Ash_Alcalinity	-0.00467	0.02645	0.93859	-0.33097	0.06435	-0.02407	0.03153	0.01515	-0.00204	0.00355	-0.05282	0.00539	0.00194
Magnesium	0.01787	0.99934	-0.02978	-0.00539	-0.00615	-0.00192	0.00180	-0.00355	0.00196	-0.00004	-0.00302	0.00062	0.00228
Total_Phenols	0.00099	0.00088	-0.04048	-0.07458	0.31525	0.27872	-0.02019	-0.17724	-0.25567	0.84720	0.00880	0.00388	-0.02669
Flavanoids	0.00157	-0.00005	-0.08544	-0.16909	0.52476	0.43360	-0.03887	-0.24812	-0.37831	-0.52014	-0.13320	-0.03749	0.06960
Nonflavanoid_Phenols	-0.00012	-0.00135	0.01351	0.01081	-0.02965	-0.02195	-0.00467	0.00650	-0.03675	0.03771	0.19918	0.14755	0.96647
Proanthocyanins	0.00060	0.00500	-0.02466	-0.05012	0.25118	0.24188	-0.30980	0.87043	0.05152	0.00972	0.13562	-0.01312	-0.01760
Color_Intensity	0.00233	0.01510	0.29140	0.87889	0.33175	0.00274	-0.11284	-0.08129	0.09903	-0.02315	-0.00982	0.05036	-0.00463
Hue	0.00017	-0.00076	-0.02598	-0.06003	0.05152	-0.02378	0.03082	-0.00295	-0.03307	-0.03847	0.09751	0.97556	-0.16655
OD280_OD315	0.00070	-0.00350	-0.07032	-0.17820	0.26064	0.28891	0.10197	-0.18671	0.87375	0.01702	0.02849	0.01163	0.04419
Proline	0.99982	-0.01777	0.00453	-0.00311	-0.00230	-0.00121	-0.00108	0.00001	0.00007	0.00005	-0.00024	-0.00010	0.00004

Principal Components: on Covariances

Covariance Matrix

	Alcohol	Malic_Acid	Ash	Ash_Alcalinity	Magnesium	Total_Phenols	Flavanoids	Nonflavanoid_Phenols	Proanthocyanins	Color_Intensity	Hue	OD280_OD315	Proline
Alcohol	0.65906	0.08561	0.04712	-0.84109	3.13988	0.14689	0.19203	-0.01575	0.06352	1.02828	-0.01331	0.04170	164.56718
Malic_Acid	0.08561	1.24802	0.05028	1.07633	-0.87078	-0.23434	-0.45863	0.04073	-0.14115	0.64484	-0.14333	-0.29245	-67.54887
Ash	0.04712	0.05028	0.07526	0.40621	1.12294	0.02215	0.03153	0.00636	0.00152	0.16465	-0.00468	0.00076	19.31974
Ash_Alcalinity	-0.84109	1.07633	0.40621	11.15269	-3.97476	-0.67115	-1.17208	0.15042	-0.37718	0.14502	-0.20912	-0.65623	-463.3553
Magnesium	3.13988	-0.87078	1.12294	-3.97476	203.98934	1.91647	2.79309	-0.45556	1.93283	6.62052	0.18085	0.66931	1769.1587
Total_Phenols	0.14689	-0.23434	0.02215	-0.67115	1.91647	0.39169	0.54047	-0.03505	0.21937	-0.08000	0.06204	0.31102	98.17106
Flavanoids	0.19203	-0.45863	0.03153	-1.17208	2.79309	0.54047	0.99772	-0.06687	0.37315	-0.39917	0.12408	0.55826	155.44749
Nonflavanoid_Phenols	-0.01575	0.04073	0.00636	0.15042	-0.45556	-0.03505	-0.06687	0.01549	-0.02606	0.04012	-0.00747	-0.04447	-12.20359
Proanthocyanins	0.06352	-0.14115	0.00152	-0.37718	1.93283	0.21937	0.37315	-0.02606	0.32759	-0.03350	0.03866	0.21093	59.55433
Color_Intensity	1.02828	0.64484	0.16465	0.14502	6.62052	-0.08000	-0.39917	0.04012	-0.03350	5.37445	-0.27651	-0.70581	230.76748
Hue	-0.01331	-0.14333	-0.00468	-0.20912	0.18085	0.06204	0.12408	-0.00747	0.03866	-0.27651	0.05224	0.09177	17.00022
OD280_OD315	0.04170	-0.29245	0.00076	-0.65623	0.66931	0.31102	0.55826	-0.04447	0.21093	-0.70581	0.09177	0.50409	69.92753
Proline	164.56718	-67.54887	19.31974	-463.3553	1769.1587	98.17106	155.44749	-12.20359	59.55433	230.76748	17.00022	69.92753	99166.717

Summary Plots



Under the Principal component 1, Proline has highest variance. It is because it has the higher scale as compared to others.

b. Comment on the use of correlations versus covariances. Would the results change dramatically if PCA (in this example) were conducted on the correlations instead?

Principal Components: on Correlations

Eigenvalues

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	4.7059	36.199		36.199
2	2.4970	19.207		55.406
3	1.4461	11.124		66.530
4	0.9190	7.069		73.599
5	0.8532	6.563		80.162
6	0.6417	4.936		85.098
7	0.5510	4.239		89.337
8	0.3485	2.681		92.018
9	0.2889	2.222		94.240
10	0.2509	1.930		96.170
11	0.2258	1.737		97.907
12	0.1688	1.298		99.205
13	0.1034	0.795		100.000

Principal Components: on Covariances

Eigenvalues

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	99201.79	99.809		99.809
2	172.5353	0.174		99.983
3	9.4381	0.009		99.992
4	4.9912	0.005		99.997
5	1.2288	0.001		99.998
6	0.8411	0.001		99.999
7	0.2790	0.000		100.000
8	0.1514	0.000		100.000
9	0.1121	0.000		100.000
10	0.0717	0.000		100.000
11	0.0376	0.000		100.000
12	0.0211	0.000		100.000
13	0.0062	0.000		100.000

Figure 4.17 Principal Components of Normalized and Nonnormalized Wine Data.

As we can notice, the covariance is not entirely distributed. The first component itself is taking 99.808%. Since, proline is measures in different weight which is more than other measures or variable present. Where as in the correlation chart , the distribution is further more than covariance, and also the variance in weights are further considered and the data is normalized too.