PCA Analysis

Mpelias Michael

25 March 2015

Contents

1 Introduction **1**

2 Principal Compontents Analysis **2**

2.1 1st Step 2

2.2 2nd 3

2.2.1 Kaiser criterion 7

2.2.2 Screeplot 9

2.2.3 Rotation 10

2.3 3rd 10

2.3.1 Screeplot 13

1 Introduction

The GreekStudyACS Example file concerns a cross-sectional study for the correlation of various risk factors with ACS (Acute coronary syndrome) and contains among others the following variables on hematology, biochemical, socio-demographic and nutritional characteristics of patients admitted to hospital with Acute Coronary Syndrome. Variables are :

- cpk mb: Levels of MB isoenzyme of CPK (ng / ml)
- troponi: Levels of troponin I (ng / ml)
- WBC: White blood cell count (number of cells / dL)
- ouria: Levels of urea (mg / dL)
- creatinine: Creatinine Levels (mg / dL)
- uric acid: Uric acid (mg / dL)
- age: Age in years
- dating only: Sex (Male: 1 & Women: 0)
- weight: Weight kgr
- height: Height in cm
- legumes: legume consumption (times / week, 0-5)
- vegetabl: Vegetable Consumption (times / week, 0-5)

- salads: salad consumption (times / week, 0-5)
- meat: meat consumption (times / week, 0-5)
- chicken: chicken consumption (times / week, 0-5)
- fish: fish consumption: (times. / week, 0-5)

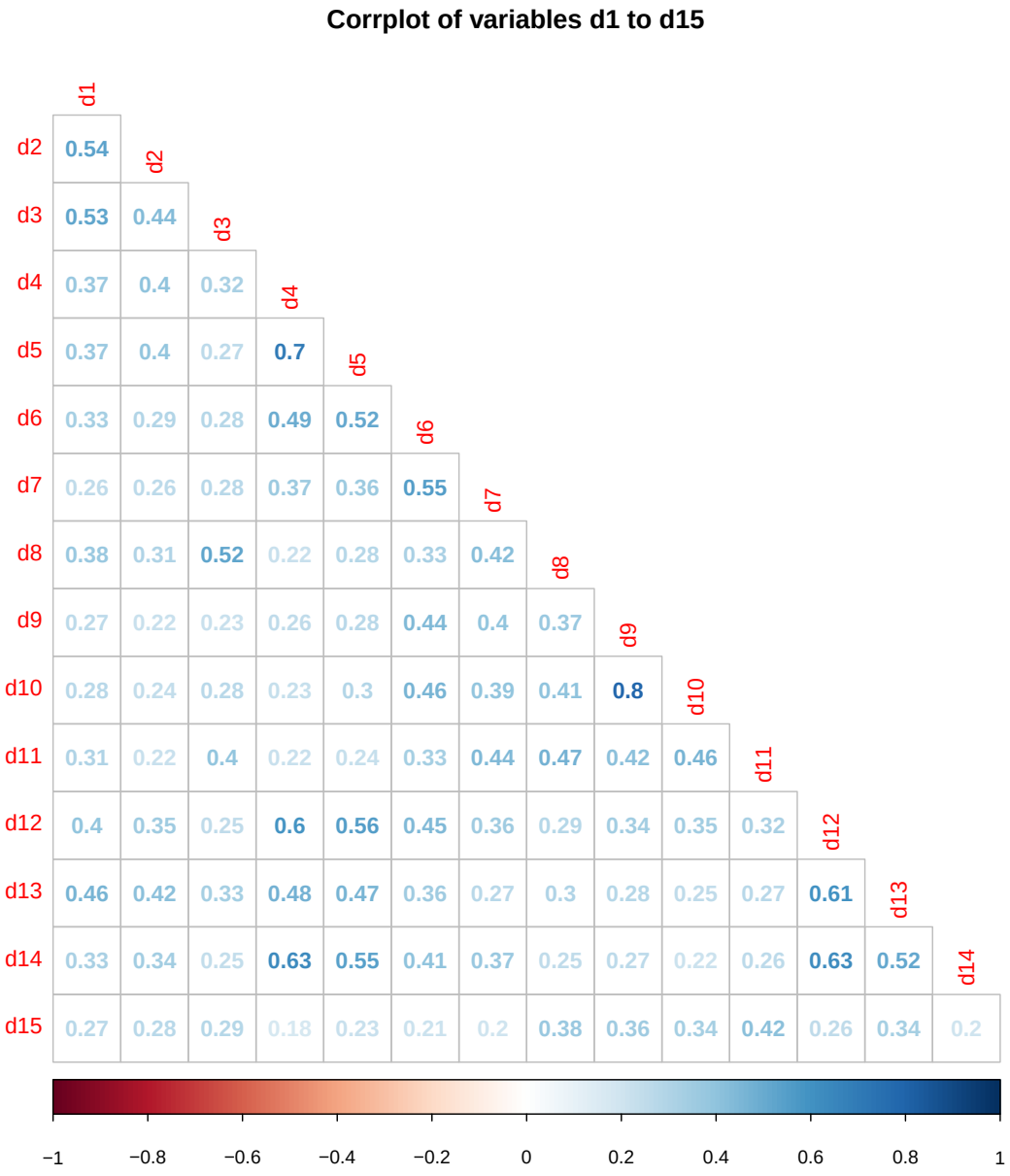
```
library(foreign)

PCA<- read.dta("C:/Users/Mike/Desktop/My Complete Book In R/Datasets/Multivariate Analys
```

2 Principal Compontents Analysis

2.1 1st Step

A prerequisite for the application of principal components analysis is that the variables to be analyzed should be continuous or bivalent. However, if the variables of our analysis (which are categorical and ordinal variables with five levels and more) can be considered to adequately converge to the normal distribution and used as such. But in all other cases we can transform the Categorical into bivalent using a cut-off of Medical importance and meaning.

```
library(corrplot)

corrplot(corr = cor(PCA[,47:61],use = "na.or.complete"),
         type = "lower",method = "number",
         title = "Corrplot of variables d1 to d15",diag = F ,mar =  c(0,0,2,0))
```



In the correlation plot we observe that there are enough correlations with $|r| > 0.3$ between variables in order to proceed to implement the method of analysis in main components.

2.2 2nd

The principal component analysis is a technique that requires a large sample size. It is suggested that observations ratio variables should be 10: 1 and that reliable results are obtained when we have a sample size of at least 300 observations. It has been suggested as a guide the following rule, according to which PCA with sample size of 50 observations class does not give good results, 300-500 comments good or very good while samples with 1000 observations give excellent results. If we have missing DATA in our sample, the default

analysis R is making is the pairwise deletion of the remarks that have at least one of the participating variables a missing value. To implement this option data should be checked for missing values, which should be either CMAR (completed missing at random) or MAR (missing at random).

In R the function for the PCA is `prcomp`

```
library(knitr)
pca_anal<-prcomp(~.,data=PCA[,c(47:61)],na.action = na.omit,scale. = T)
```

The attributes inhereted in this analysis are the following:

```
a=summary(pca_anal)

eig <- (pca_anal$sdev)^2

#Difference to the next EigenValue
difference <- vector()
for(i in 1:length(eig)){difference[i] <- eig[i] - eig[i+1]}



# Variances in percentage
Proportion <- eig/sum(eig)

# Cumulative variances
cumvar <- cumsum(Proportion)

PCA.eigen <- data.frame(Eigenvalue = (pca_anal$sdev)^2, Difference = difference, Proport
                        Cumulative = cumvar)


kable( PCA.eigen  ,caption = "Principal Component/Correlation")
```

Table 1: Principal Component/Correlation

Eigenvalue	Difference	Proportion	Cumulative
6.0923987	4.3298317	0.4061599	0.4061599
1.7625670	0.4326541	0.1175045	0.5236644
1.3299129	0.4315971	0.0886609	0.6123252
0.8983158	0.0937828	0.0598877	0.6722130
0.8045330	0.1880539	0.0536355	0.7258485
0.6164792	0.0527081	0.0410986	0.7669471
0.5637711	0.0512328	0.0375847	0.8045318

Eigenvalue	Difference	Proportion	Cumulative
0.5125383	0.0175591	0.0341692	0.8387011
0.4949792	0.0823031	0.0329986	0.8716997
0.4126762	0.0188857	0.0275117	0.8992114
0.3937905	0.0450726	0.0262527	0.9254641
0.3487179	0.0278322	0.0232479	0.9487120
0.3208857	0.0591012	0.0213924	0.9701044
0.2617845	0.0751344	0.0174523	0.9875567
0.1866501	NA	0.0124433	1.0000000

In the first table are the main components derived from the PCA, the eigenvalue for each main component, the difference between the ith and the next main component, the percentage of the initial fluctuation interpreting each main component and the corresponding cumulative percentage of the initial variation is interpreted.

```
library(pander)

un = vector()
for(i in 1:length(PCA.eigen[,1])){

un[i] = round(1- sum((a$rotation[i,1:15]^2)* PCA.eigen$Eigenvalue),4)

}

pander(cbind(round(a$rotation[1:length(a$sdev), 1:length(a$sdev)] , 4) , Unexplained=un
        caption = "Principal Components (Eigenvectors)", split.table = 75 )
```

Table 2: Principal Components (Eigenvectors) (continued below)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
d1	0.2577	-0.0448	0.407	-0.0386	0.3253	-0.1622	0.1901	0.024
d2	0.2409	-0.1153	0.3679	-0.0161	0.3841	0.3272	0.3217	0.2623
d3	0.2353	0.1174	0.4756	-0.2779	0.0246	-0.1552	-0.362	0.0303
d4	0.283	-0.3609	-0.1161	-0.0809	-0.0088	0.1499	-0.3384	0.1882
d5	0.2838	-0.2957	-0.1349	-0.0394	0.031	0.3252	-0.3746	-0.0055
d6	0.2783	-0.0018	-0.2928	-0.3098	0.1251	0.2437	0.1287	-0.3007

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
d7	0.2523	0.123	-0.2326	-0.4858	-0.2202	0.1202	0.4944	-0.0422
d8	0.2463	0.2849	0.2195	-0.2129	-0.2582	-0.1218	-0.2742	-0.4802
d9	0.2488	0.3574	-0.3092	0.2229	0.3496	-0.1135	-0.0812	0.0193
d10	0.2516	0.3861	-0.2737	0.1549	0.3662	-0.1164	-0.143	0.0628
d11	0.2398	0.3341	0.0302	-0.0285	-0.4193	-0.1404	0.0844	0.6091
d12	0.2942	-0.2459	-0.1373	0.2263	-0.1197	-0.3303	0.0825	-0.0156
d13	0.2739	-0.2144	0.117	0.3478	-0.0688	-0.2816	0.3095	-0.3869
d14	0.2711	-0.3204	-0.1347	0.0676	-0.2293	-0.2066	-0.0336	0.1849
d15	0.2001	0.2438	0.1759	0.5333	-0.3323	0.5903	0.0189	-0.1095

	PC9	PC10	PC11	PC12	PC13	PC14	PC15	Unexplained
d1	-0.3823	0.1212	-0.6444	-0.1029	-0.0734	0.0434	0.0398	0
d2	0.476	0.1234	0.256	0.2353	0.0086	0.0378	-0.0436	0
d3	-0.1472	-0.4783	0.3137	-0.0805	0.2081	-0.2599	-0.068	0
d4	-0.0309	-0.0961	0.0145	-0.2292	0.1138	0.7105	0.1281	0
d5	-0.0745	0.4461	0.0156	-0.2482	-0.1395	-0.5129	-0.1277	0
d6	-0.4345	-0.1303	0.1386	0.5511	-0.1553	0.0668	-0.0401	0
d7	0.1689	-0.1163	-0.1009	-0.4564	0.2168	-0.1098	0.0488	0
d8	0.4033	0.3321	-0.1479	0.1697	-0.0912	0.2132	-0.0057	0
d9	0.1345	-0.1606	-0.0611	-0.1618	-0.076	0.1057	-0.6575	0
d10	0.0676	0.0397	0.0662	-0.0226	0.0499	-0.1212	0.6989	0

	PC9	PC10	PC11	PC12	PC13	PC14	PC15	Unexplained
d11	-0.2555	0.2987	0.1827	0.0972	-0.215	0.0832	-0.0706	0
d12	-0.0143	0.1799	-0.0188	0.2919	0.7147	-0.1027	-0.1144	0
d13	-0.1061	0.0206	0.4508	-0.3078	-0.308	0.076	0.0675	0
d14	0.3329	-0.4358	-0.3062	0.2379	-0.4011	-0.2233	0.093	0
d15	-0.0942	-0.2303	-0.1739	0.0029	0.1355	-0.0197	0.0558	0

The second table refers to the main elements that resulted from the analysis and the elements of eigenvector that constituent them.

With principal component analysis 15 new elements will created as many as the original variables introduced in the analysis, being a mathematical transformation of the original variables.

2.2.1 Kaiser criterion

The Kaiser criterion proposes to maintain for further analysis of heart rate the variables with eigenvalue greater than 1.

The variables that satisfy the Kaiser criterion are the first 3 main components

```
count=length(PCA.eigen[which(PCA.eigen$Eigenvalue >=1),]$Eigenvalue)

kable(PCA.eigen[which(PCA.eigen$Eigenvalue >=1),],
      caption = "Principal Components (Kaiser criterion)")
```

Table 4: Principal Components (Kaiser criterion)

Eigenvalue	Difference	Proportion	Cumulative
6.092399	4.3298317	0.4061599	0.4061599
1.762567	0.4326541	0.1175045	0.5236644
1.329913	0.4315971	0.0886609	0.6123252

```
un = vector()
for(i in 1:length(PCA.eigen[,1])){
```

```
un[i] = round(sum((a$rotation[i,4:15]^2)* PCA.eigen$Eigenvalue[4:15]) ,4)

}

pander(cbind(round(a$rotation[1:length(a$sdev), 1:count] , 4) , Unexplained=un),
        caption = "Principal Components (Eigenvectors)", split.table = 75 )
```

Table 5: Principal Components (Eigenvectors)

	PC1	PC2	PC3	Unexplained
d1	0.2577	-0.0448	0.407	0.3715
d2	0.2409	-0.1153	0.3679	0.4429
d3	0.2353	0.1174	0.4756	0.3376
d4	0.283	-0.3609	-0.1161	0.2646
d5	0.2838	-0.2957	-0.1349	0.3311
d6	0.2783	-0.0018	-0.2928	0.4142
d7	0.2523	0.123	-0.2326	0.5137
d8	0.2463	0.2849	0.2195	0.4233
d9	0.2488	0.3574	-0.3092	0.2704
d10	0.2516	0.3861	-0.2737	0.252
d11	0.2398	0.3341	0.0302	0.4517
d12	0.2942	-0.2459	-0.1373	0.3411
d13	0.2739	-0.2144	0.117	0.4437
d14	0.2711	-0.3204	-0.1347	0.3471
d15	0.2001	0.2438	0.1759	0.6102

In the third column of Table 4 the percentage of the initial variability expressed by each main component is reported. The percentage of variance is calculated with the formula $\frac{\lambda_i}{\sum \lambda_i}$ where λ_i the eigenvalue of each variable and $\sum \lambda_i$ the sum of the eigenvalues which because it was used in the correlation table equals the number of main components and the number initial variables respectively.

For the first component, the rate of the total variance that interprets equals 0.406 of the total variance, which is confirmed by the figure. Similarly the proportion of variance that interprets the second heart rate is 0.118 of the total variance and the third CS the figure is 0.089.

The eigenvectors corresponding to the Main Components eigenvalues are described in Table 2. The data of the eigenvectors are the coefficients of the variables, the linear combination of which constitutes each main component and they express the correlation of each variable with the corresponding principal component. Because the values of the eigenvectors have great uniformity with their Main Componetn values , it is very difficult to characterize each Main Component of the variables in which the elements have the greatest value. To improve the interpretation, the results should be rotated. The availiable rotations may be

rectangular (Main Components rotated are uncorrelated to each other) or non-rectangular (Main Components rotated are correlated to each other).

The value of each main component is the sum of the product of elements of the eigenvectors of each variable with the corresponding value of each variable. Using the values of the initial eigenvectors (Table 2) the value of the first principal component is

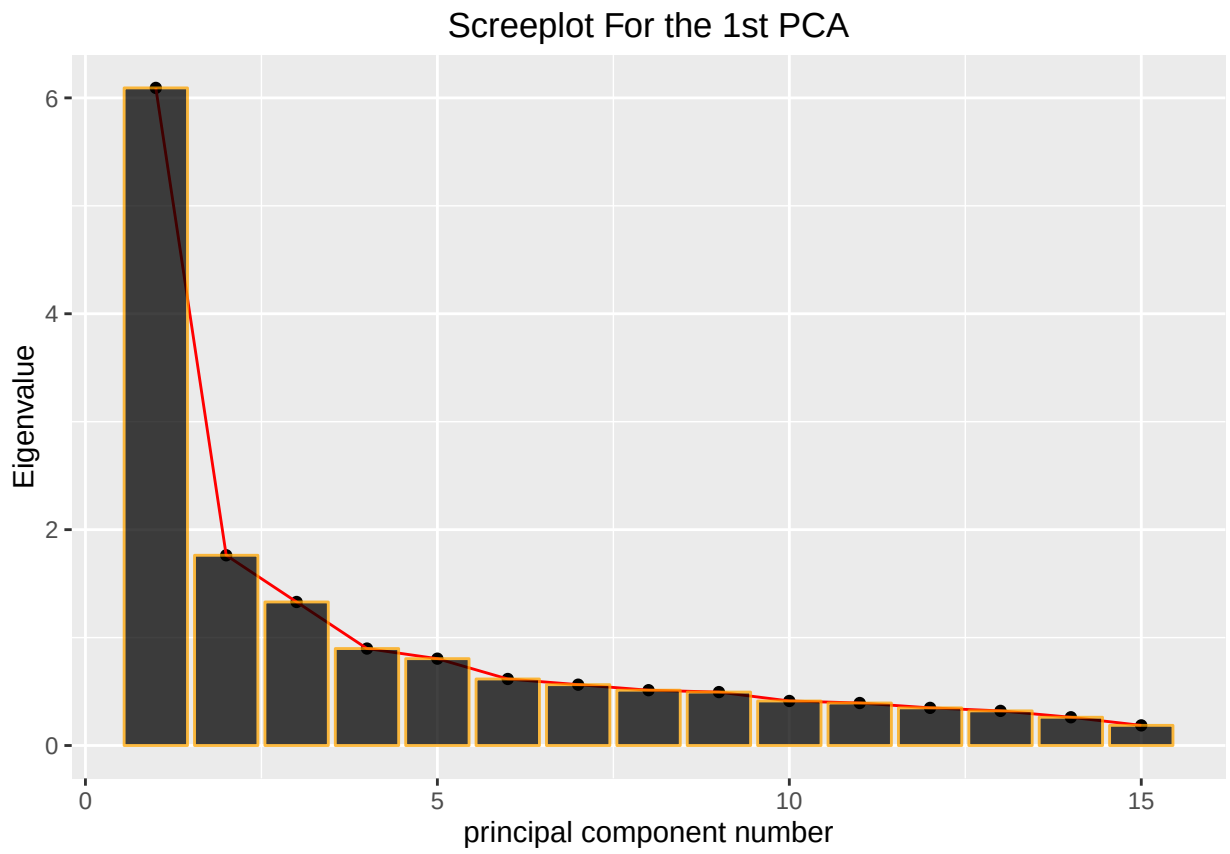
$Y_1 = 0.258 * X_1 + 0.241 * X_2 + 0.235 * X_3 + 0.283 * X_4 + 0.284 * X_5 + 0.278 * X_6 + 0.252 * X_7 + 0.246 * X_8 + 0.249 * X_9 + 0.252 * X_10 + 0.24 * X_11 + 0.294 * X_12 + 0.274 * X_13 + 0.271 * X_14 + 0.2 * X_15$

The main components generated as a linear combination of the original variables of the analysis do not have units.

2.2.2 Screeplot

```
ggscreeplot <- function(pca_odject, type = c('norm','pev', 'cev'))
{
  require(ggplot2)
  type <- match.arg(type)
  d <- pca_odject$sdev^2
  yvar <- switch(type,
                 norm= d,
                 pev = d / sum(d),
                 cev = cumsum(d) / sum(d))

  yvar.lab <- switch(type,
                     norm= "Eigenvalue",
                     pev = 'proportion of explained variance',
                     cev = 'cumulative proportion of explained variance')

  df <- data.frame(PC = 1:length(d), yvar = yvar)

  ggplot(data = df, aes(x = PC, y = yvar)) +
    xlab('principal component number') + ylab(yvar.lab) +
    geom_point() + geom_path(colour="red") + geom_bar(stat= "identity",fill="black",col
}

ggscreeplot(pca_anal,"norm") + ggtitle("Screeplot For the 1st PCA")
```

```
## Loading required package: ggplot2
```



2.2.3 Rotation

2.3 3rd

With the principal component analysis the use of Variance - Covariance matrix is sensitive to the change of scale of measurement units of the variables and the variations of magnitude as the variable that has the largest variance value tends to be identified as the first principal component. The application of PCA using the Variance - Covariance matrix only makes sense if we want to maintain in our analysis the differences in the variability of the initial variables. We should apply PCA using similar methodology but instead of Variance - Covariance matrix we will use the correlation table by adding the `scale. = F` option in the `prcomp` command.

Table 6: Principal Component/Covariance

Eigenvalue	Difference	Proportion	Cumulative
7.1228622	5.0627788	0.4011532	0.4011532
2.0600834	0.3919778	0.1160220	0.5171752
1.6681056	0.5938220	0.0939462	0.6111214
1.0742836	0.0655387	0.0605027	0.6716241
1.0087449	0.2652029	0.0568116	0.7284357

Eigenvalue	Difference	Proportion	Cumulative
0.7435421	0.0646264	0.0418756	0.7703113
0.6789157	0.0539215	0.0382359	0.8085473
0.6249942	0.0346201	0.0351991	0.8437464
0.5903741	0.0871206	0.0332493	0.8769957
0.5032535	0.0523173	0.0283428	0.9053385
0.4509362	0.0716666	0.0253963	0.9307348
0.3792695	0.0408552	0.0213601	0.9520949
0.3384144	0.0529184	0.0190592	0.9711541
0.2854959	0.0588053	0.0160789	0.9872330
0.2266907	NA	0.0127670	1.0000000

And the Eigen Vector Matrix is:

Table 7: Principal Components(Eigenvectors) (continued below)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
d1	0.2708	0.1113	0.3643	-0.0203	0.2948	-0.1461	0.2798	-0.0408
d2	0.2646	0.199	0.3477	-0.1123	0.4058	0.5401	0.1819	-0.1285
d3	0.2723	-0.0693	0.5228	0.294	0.0276	-0.3013	-0.2075	0.374
d4	0.2594	0.3603	-0.1518	0.0592	-0.0551	-0.0204	-0.1757	0.335
d5	0.2577	0.2938	-0.1651	0.0074	-0.027	0.0863	-0.2905	0.2691
d6	0.2589	0.026	-0.2666	0.263	0.0947	0.2302	-0.1257	0.1348
d7	0.248	-0.0958	-0.2091	0.4603	-0.1559	0.4355	0.1755	-0.2458
d8	0.2701	-0.2653	0.2306	0.245	-0.2302	-0.0984	-0.4519	-0.5276
d9	0.2589	-0.3529	-0.3093	-0.172	0.3677	-0.1473	-0.0523	-0.0256
d10	0.2658	-0.3829	-0.2703	-0.1054	0.3885	-0.1587	-0.0408	0.0526
d11	0.2651	-0.3437	0.0322	0.0709	-0.4119	-0.0538	0.6076	0.3094
d12	0.264	0.2364	-0.1794	-0.1503	-0.1247	-0.2677	0.1476	-0.189
d13	0.2488	0.2201	0.0338	-0.2814	-0.0874	-0.2268	0.1198	-0.3635

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
d14	0.2451	0.3081	-0.1729	-0.0309	-0.2112	-0.1583	0.0604	-0.1062
d15	0.2185	-0.2414	0.1473	-0.6383	-0.3661	0.3717	-0.2571	0.1486

	PC9	PC10	PC11	PC12	PC13	PC14	PC15	Unexplained
d1	0.5225	-0.375	0.4202	-0.0145	-0.03	-0.0112	0.0387	-0.2399
d2	-0.4501	0.0298	-0.1646	0.1303	0.0333	-0.0273	-0.0401	-0.3507
d3	-0.021	0.4862	-0.0691	-0.0569	0.1106	0.156	-0.0734	-0.4068
d4	-0.1418	-0.0598	0.1327	-0.1897	0.0053	-0.7173	0.1949	-0.0741
d5	-0.0625	-0.4403	-0.1349	-0.3302	-0.1074	0.5334	-0.1841	-0.0473
d6	0.4643	0.0401	-0.3451	0.5887	-0.089	-0.0614	-0.0397	-0.0595
d7	0.1583	0.2778	0.2322	-0.4338	0.1244	0.0819	0.0451	-0.1386
d8	-0.1756	-0.3598	-0.0206	0.1174	-0.0394	-0.149	0.0029	-0.2866
d9	-0.0854	0.1302	0.1714	-0.0943	-0.1613	-0.1673	-0.6375	-0.2185
d10	-0.1265	-0.023	-0.051	-0.0459	0.1024	0.1753	0.6749	-0.2403
d11	-0.1992	-0.2514	-0.1892	0.0603	-0.1379	-0.0566	-0.0589	-0.3158
d12	-0.016	-0.0016	-0.1115	0.1185	0.7864	0.0443	-0.1669	-0.0229
d13	0.2158	0.2513	-0.4873	-0.2991	-0.3913	-0.0669	0.0976	-0.0214
d14	-0.2649	0.2425	0.4853	0.4088	-0.3356	0.2732	0.1035	-0.0499
d15	0.2203	0.1174	0.1708	0.0285	0.098	0.0089	0.0503	-0.2835

In PCA using the Variance - Covariance matrix 15 main Components will be created , as many as the primary variables for analysis. Implementing the Kaiser criterion we shall keep the main components that have eigenvalue greater than the average value of the eigenvalues of the analysis, that is $\frac{\sum E_i}{n}$= 1.1837311 - where E_i : each Eigen Value and n the amount (in

our case 15 - so the first 3 principal components will be maintained.

```
kable(PCA.eigen1[which(PCA.eigen1$Eigenvalue >=Kaiser.Cut),],
      caption = "Principal Components (Kaiser criterion)")
```

Table 9: Principal Components (Kaiser criterion)

Eigenvalue	Difference	Proportion	Cumulative
7.122862	5.0627788	0.4011532	0.4011532
2.060083	0.3919778	0.1160220	0.5171752
1.668106	0.5938220	0.0939462	0.6111214

2.3.1 Screeplot

```
ggscreeplot(pca_anal2,"norm") + ggtitle("Screeplot For the 2nd PCA")
```