Principal Component Analysis

T Shapla

10/25/2023

## Principal Component Analysis (PCA)

## Example: Places Rated Data

In the Places Rated Almanac, Boyer and Savageau rated 329 communities according to the following nine criteria:

1. Climate and Terrain
2. Housing
3. Health Care and Environment
4. Crime
5. Transportation
6. Education
7. Arts
8. Recreation
9. Economics

There are 329 observations representing the 329 communities in the dataset and 9 variables. First we draw histograms for each variable to display the data distribution.

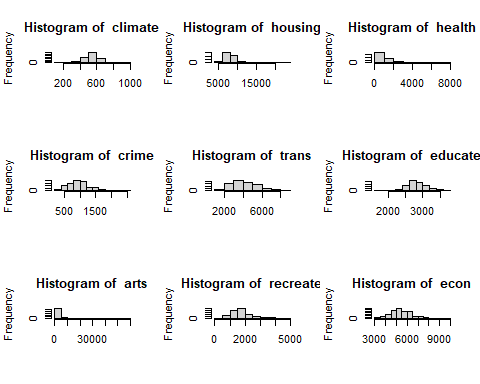
setwd("D:/LianZuo/Applied Statistics Course Materials/STAT 577 - Applied Multivariate Statistics/data-577")  
places=read.csv("places.csv", header=T)  
head(places, 3)

climate housing health crime trans educate arts recreate econ id  
1 521 6200 237 923 4031 2757 996 1405 7633 1  
2 575 8138 1656 886 4883 2438 5564 2632 4350 2  
3 468 7339 618 970 2531 2560 237 859 5250 3

names(places)

[1] "climate" "housing" "health" "crime" "trans" "educate"   
 [7] "arts" "recreate" "econ" "id"

#Drawing histograms  
par(mfrow=c(3,3))  
hst=function(x){  
 y=places[,x]  
 hist(y, main=paste("Histogram of ", x), xlab="")  
}  
invisible(sapply(names(places[1:9]), hst))



The data for many of the variables appear to be strongly skewed to the right. So, we take the log transformation of all the variables (except the last one which is the id variable) to normalize the data. However, note that normality is not required to find the principal components.

Note: We take logarithmic transformation on variables whose domain is the set of positive real numbers. For a right skewed distribution, the log transformation compresses larger values more than smaller ones. This helps reducing the skewness of the data and making it more symmetric.

places.trs=matrix(0, 329, 9) #creating a matrix of zeros  
  
for (i in 1:9){  
 places.trs[,i]=log10(places[,i])  
 }  
colnames(places.trs)=c("climate", "housing", "health","crime","trans", "educate", "arts", "recreate", "econ")  
places.trs=data.frame(places.trs)  
head(places.trs, 3)

climate housing health crime trans educate arts recreate  
1 2.716838 3.792392 2.374748 2.965202 3.605413 3.440437 2.998259 3.147676  
2 2.759668 3.910518 3.219060 2.947434 3.688687 3.387034 3.745387 3.420286  
3 2.670246 3.865637 2.790988 2.986772 3.403292 3.408240 2.374748 2.933993  
 econ  
1 3.882695  
2 3.638489  
3 3.720159

Below we find the eigenvalues of the covariance matrix, and the sum of the eigenvalues which is equal to the trace of sample covariance matrix, and is called the “total variance” of underlying variables.

s.cov=cov(places.trs)#finding covariance matrix  
lambdas=round(eigen(s.cov)$values, digits=4)#eigenvalues of cov matrix  
lambdas

[1] 0.3775 0.0511 0.0279 0.0230 0.0168 0.0120 0.0085 0.0039 0.0018

sum(lambdas)#sum of eigenvalues

[1] 0.5225

sum(diag(s.cov))#trace of sample of sample covariance matrix

[1] 0.5223134

Now we perform the PCA for Places Rated data using princomp() function from basic “stats” package. Since all the variables are measure in score, we use sample covariance matrix instead of correlation matrix to find the loadings for principal components. The option cor=F is used in princomp() function to specify the use of sample covariance matrix. Note that a sample correlation matrix is used for principal component analysis when the variables are measured in different scales or units.

pc.places<-princomp(places.trs, cor=F)#using sample covariance matrix  
result.places=summary(pc.places)  
result.places

Importance of components:  
 Comp.1 Comp.2 Comp.3 Comp.4 Comp.5  
Standard deviation 0.6134452 0.22560372 0.1668374 0.15131845 0.12930691  
Proportion of Variance 0.7226740 0.09774248 0.0534537 0.04397184 0.03210956  
Cumulative Proportion 0.7226740 0.82041652 0.8738702 0.91784205 0.94995161  
 Comp.6 Comp.7 Comp.8 Comp.9  
Standard deviation 0.10916208 0.09182048 0.062627954 0.042330504  
Proportion of Variance 0.02288413 0.01619086 0.007532295 0.003441103  
Cumulative Proportion 0.97283574 0.98902660 0.996558897 1.000000000

The proportion of variation explained by each eigenvalue is given in the second row labeled by “Proportion of Variance”. For example, = 0.3775 divided by = 0.5223 equals 0.7227, or, about 72% of the variation in the data is explained by this first principal component. The cumulative proportion is obtained by adding the successive proportions of variation to obtain the running total. For instance, 0.7227 plus 0.0977 equals 0.8204, and so forth. Therefore, about 82% of the variation is explained by the first two principal components together.

The first row labeled by “Standard deviation” provides the standard deviation of each principal component, which is equal to the square root of the eigenvalues of the covariance matrix.

,

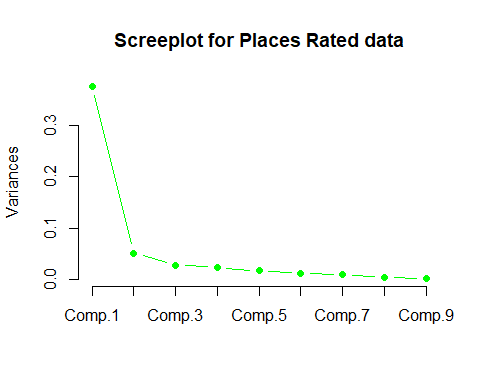
Next, we need to look at successive differences between the eigenvalues.

For the Places Rated example, we find as follows ordered from largest to smallest:

0.3775, 0.0511, 0.0279, 0.0230, 0.0168, 0.0120, 0.0085, 0.0039, 0.0018

Subtracting the second eigenvalue 0.0511 from the first eigenvalue, 0.3775 we get a difference of 0.3264. The difference between the second and third eigenvalues is 0.0232; the next difference is 0.0049. Subsequent differences are even smaller. A sharp drop from one eigenvalue to the next may serve as another indicator of how many eigenvalues to consider. A scree plot is the plot of versus , which is a graphical way to determine the number of principal components to retain. The number of components is determined at the point beyond which the remaining eigenvalues are all relatively small and of comparable size. Below we draw the scree plot using screeplot() functio from basic “stats” package. In screeplot() function, the first argument is the object from princomp() function.

screeplot(pc.places, pch=16, col="green", cex=1, type = "lines",   
 lwd=1.5, main = "Screeplot for Places Rated data")



In the screeplot, an elbow occurs in the plot at . That is, the eigenvalues after are all relatively small and about the same size. In this case, it appears that two or perhaps three sample principal components effectively summarize the total sample variance. First three principal components collectively explain 87% of the total variation. This is an acceptably large percentage. Consequently, sample variation is summarized very well by first three principal components and a reduction in the data from 329 observations on 9 variables to 329 observations on 3 principal components is reasonable.

Next, we will write down the estimated principal components. For Places Rated example, the first principal component can be written using the loadings or scores as provided in the R output below. In loadings, omitted values are not zero but are close to zero in value and should be ignored. In R results, loadings with an absolute value less than 0.1 are not displayed.

result.places$loadings

Loadings:  
 Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9  
climate 0.141 0.153 0.398 0.831 0.315   
housing 0.129 -0.178 0.175 0.209 -0.696 -0.614   
health 0.408 -0.859 0.276   
crime 0.100 0.220 0.593 0.724 -0.164 -0.182   
trans 0.150 0.221 -0.126 -0.870 0.372   
educate 0.204 -0.973  
arts 0.874 0.304 -0.363   
recreate 0.159 0.334 0.584 -0.628 0.213 -0.142 0.235   
econ 0.121 -0.265 -0.664 0.662 0.168  
  
 Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9  
SS loadings 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000  
Proportion Var 0.111 0.111 0.111 0.111 0.111 0.111 0.111 0.111 0.111  
Cumulative Var 0.111 0.222 0.333 0.444 0.556 0.667 0.778 0.889 1.000

The estimated first three principal components can be computed using the elements appear columns Comp.1, Comp.2, and Comp.3 in “Loadings” section of the output above, Or, you may get them using the code below:

loads.pc123=round(result.places$loadings[,1:3], digits=4)  
loads.pc123

Comp.1 Comp.2 Comp.3  
climate 0.0351 0.0089 0.1409  
housing 0.0934 0.0092 0.1288  
health 0.4078 -0.8585 0.2761  
crime 0.1004 0.2204 0.5927  
trans 0.1501 0.0592 0.2209  
educate 0.0322 -0.0606 0.0081  
arts 0.8743 0.3038 -0.3633  
recreate 0.1590 0.3340 0.5836  
econ 0.0195 0.0561 0.1209

Thus, considering the absolute value of laodings greater than 0.1, the estimated first principal component for the Places Rated data is:

=

In order to complete this formula and compute the principal component for the individual community of interest, plug in that community’s log transformed values for each of these variables. A fairly standard procedure is to use the difference between the variables and their sample means rather than the raw data. This is known as a translation of the random variables. Translation does not affect the interpretations because the variances of the original variables are the same as those of the translated variables.

The sign of the loadings or coefficients associated with a principal component indicates the direction of linear relationship with that PC and the corresponding X variable. For example, all the loadings of first PC are positive indicating that all X variables are positively correlated with first PC. For second PC, all variables are positively associated except for health and education.

The magnitudes of the coefficients give the contributions of each variable to that component. However, the magnitude of the coefficients also depends on the variances of the corresponding variables. We find the sample variance for each variable, and compare it with the associated loadings of each principal component.

round(sapply(places.trs, var), digits=4)#finding sample variance of each variable

climate housing health crime trans educate arts recreate   
 0.0129 0.0111 0.1027 0.0286 0.0248 0.0025 0.2972 0.0353   
 econ   
 0.0071

The large absolute coefficient in and , 0.874 and 0.859, respectively correspond to the two largest variances associated with arts and health variables, (0.2972 and 0.1027), respectively. The two variables with large variances have notable influences on the first two principal components. In general, variables with the highest sample variances will tend to be emphasized in the first few principal components.

We will now interpret the principal component results:

First Principal Component - PC1

The first principal component increases with increasing Arts and Health, and to some extent Transportation, and Recreation scores. This component can be viewed as a measure of the quality of Arts, Health, Transportation, and Recreation. Furthermore, we see that the highest loading in the first principal component is associated with the Arts variable. Thus, we can state that this principal component is primarily a measure of the Arts. It would follow that communities with high values tend to have a lot of arts available, in terms of theaters, orchestras, etc. Whereas communities with small values would have very few of these types of opportunities.

Second Principal Component - PC2

The second principal component increases with only one of the values, decreasing Health. This component can be viewed as a measure of how unhealthy the location is in terms of available health care including doctors, hospitals, etc.

Third Principal Component - PC3

The third principal component increases mostly with increasing Crime and Recreation. This suggests that places with high crime also tend to have better recreation facilities.