Why log-transforming the data before performing principal component analysis?

Im following a tutorial here: http://www.r-bloggers.com/computing-and-visualizing-pca-in-r/ to gain a better understanding of PCA.

The tutorial uses the Iris dataset and applies a log transform prior to PCA:

> Notice that in the following code we apply a log transformation to the continuous variables as suggested by [1] and set `center` and `scale` equal to `TRUE` in the call to `prcomp` to standardize the variables prior to the application of PCA.

Could somebody explain to me in plain English why you first use the log function on the the first four columns of the Iris dataset. I understand it has something to do with making data relative but am confused what's exactly the function of log, center and scale.

r pca

edited Aug 2 '15 at 22:31
amoeba
24.5k 5 84 143

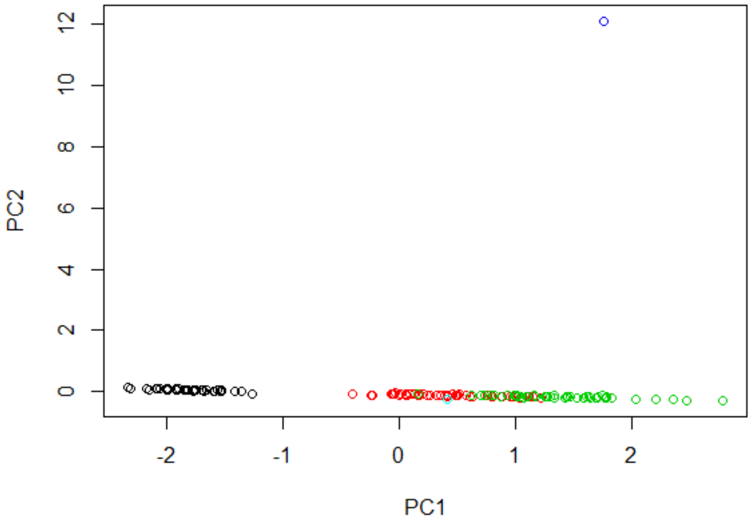
asked Aug 2 '15 at 16:05
Marc van der Peet
48 3

2 Answers

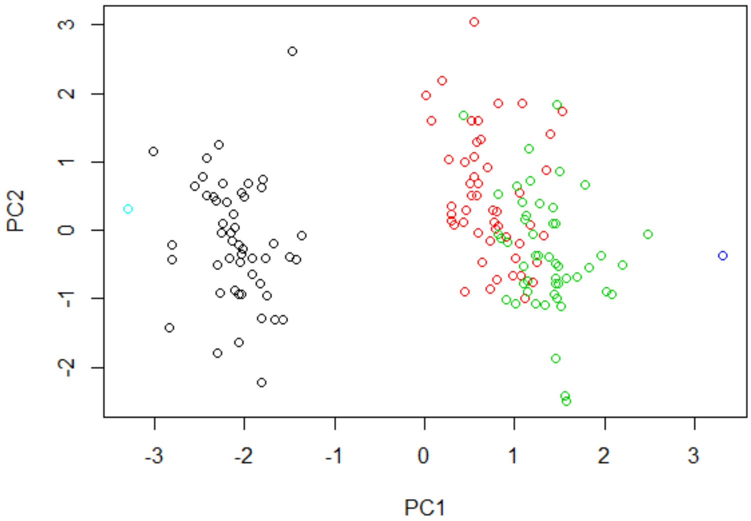
The Iris data set is a fine example to learn PCA. That said, the first four columns describing length and width of two leaf types is not an example of strongly skewed data. Therefore log-transforming the data does not change the results much, since the resulting rotation of the principle components is quite unchanged by log-transformation. In other situations log-transformation is a good choice. We perform PCA to get insight of the general structure of a data set. We center, scale and some times log-transform to filter off some trivial effects, which could dominate our PCA. The algorithm of a PCA analysis will in turns find the rotation of each PC to minimize the squared residuals. That is the sum of squared perpendicular distances from any sample to the PC. Large values tend to have high leverage. Imagine injecting two new samples into the iris data. A flower with 430.0cm petal length and one with petal length of 0.0043cm. Both flowers are very abnormal being 100 times larger and 1000 times smaller respectively than average examples. The leverage of the first flower is huge, such that the first PCs mostly will describe the differences between the large flower and any other flower. Clustering of species is not possible due to that one outlier sample. If the data are log-transformed, the absolute value now describes the relative variation. Now the small flower is the most abnormal one. Nonetheless it is possible to both contain all samples in one image and provide a fair clustering of the species. Check out this example:

```
data(iris) #get data
#add two new sample of two new species to iris data
levels(iris[,5]) = c(levels(iris[,5]),"setosa_gigantica","virginica_brevis")
iris[151,] = list(6,3,  430  ,1.5,"setosa_gigantica") #a big flower
iris[152,] = list(6,3,.0043,1.5  ,"virginica_brevis") #a small flower

#Plotting scores of PC1 and PC" without log transform
plot(prcomp(iris[,-5],cen=T,sca=T)$x[,1:2],col=iris$Spec)
```



```
#Plotting scores of PC1 and PC" without log transform
plot(prcomp(log(iris[,-5]),cen=T,sca=T)$x[,1:2],col=iris$Spec)
```



edited Jul 1 at 9:18
HelloWorld
26 6

answered Aug 2 '15 at 17:25
Soren Havelund Welling
2,871 5 17

2 Nice demo and plots. – ssdecontrol Aug 3 '15 at 0:55

Well, the other answer gives an example, when the log-transform is used to reduce the influence of extreme values or outliers.
Another general argument occurs, when you try to analyze data which are *multiplicatively* composed instead of *addititively* - PCA and FA model by their math such additive compositions. *Multiplicative* compositions occur in the most simple case in physical data like the surface and the volume of bodies (functionally) dependent on (for instance) the three parameters lenght, width, depth. One can reproduce the compositions of an historic example of the early PCA, I think it is called "Thurstone's Ball- (or 'Cubes'-) problem" or the like. Once I had played with the data of that example and had found that the log-transformed data gave a much nicer and clearer model for the composition of the measured volume and surface data with the three one-dimensional measures.

Besides of such simple examples, if we consider in social research data *interactions* , then we ususally think them as well as multiplicatively composed measurements of more elementary items. So if we look specifically at interactions, a log-transform might be a special helpful tool to get a mathematical model for the de-composition.

answered Jul 2 at 8:48
Gottfried Helms
793 6 13

Add Another Answer