PCA

KEY

Tue Oct 1 10:43:21 2024

## Principal Components Analysis of buffalo teeth

We will use PCA on the buffalo teeth data to help us understand the morphology of these complex structures.

First, import the tooth data:

library(readxl)  
data.frame(read\_excel("teeth\_fixed.xlsx", "teeth\_fixed")) -> teeth

Assign the tooth ID numbers as row labels so that they will be used to label points in our biplots:

row.names(teeth) <- teeth$tooth.num

Make your variable.names object:

variable.names <- c("weight","length", "width.crown","width.root","depth","surface.len","thickness")

Check that the correlation matrix should be expected to give us a good PCA - there should be several correlations that are above 0.4 for us to be justified in running a PCA.

cor(teeth[variable.names])

## weight length width.crown width.root depth surface.len thickness  
## weight 1.000 0.8390 0.4692 0.2080 0.4138 0.8309 0.5109  
## length 0.839 1.0000 0.4736 0.0602 0.5300 0.9482 0.2787  
## width.crown 0.469 0.4736 1.0000 -0.0641 0.0769 0.4532 0.1798  
## width.root 0.208 0.0602 -0.0641 1.0000 0.1093 0.0512 -0.0316  
## depth 0.414 0.5300 0.0769 0.1093 1.0000 0.6300 -0.1221  
## surface.len 0.831 0.9482 0.4532 0.0512 0.6300 1.0000 0.2763  
## thickness 0.511 0.2787 0.1798 -0.0316 -0.1221 0.2763 1.0000

#### Question: are there sufficient correlations between the variables to justify a PCA? Which variables are most highly correlated?

Yes, inside.surface.len and right.len are correlated at 0.822, and length and inside.surface.len are correlated at 0.792. These are large enough to justify PCA, even with a small sample size.

### Run the PCA, initial assessment of effectiveness

Run the principal components analysis:

princomp(teeth[variable.names], cor = T) -> teeth.pca  
teeth.pca

## Call:  
## princomp(x = teeth[variable.names], cor = T)  
##   
## Standard deviations:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7   
## 1.877 1.114 1.007 0.868 0.558 0.342 0.204   
##   
## 7 variables and 100 observations.

#### Question: based on the standard deviations, if we were using the Kaiser-Guttman rule for the number of components to retain and interpret, which componenents should we retain?

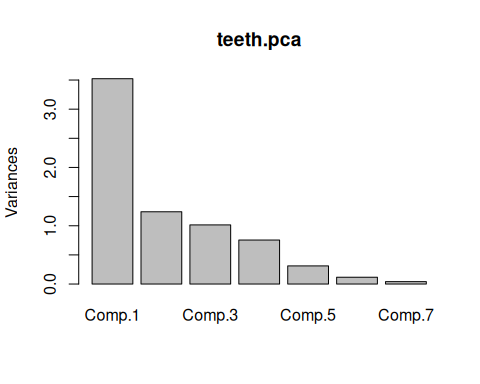
We would retain the first three components, since K-G tells us to retain compenents with variances greater than 1.

#### Question: Comp.1 has the largest standard deviation, is this a coincidence? Or will the first component always explain the most variation?

It is not a coincidence, Component 1 is selected specifically to have the largest explained variance (and thus, the largest standard deviation as well).

Plot a scree plot of the results:

plot(teeth.pca)



#### Question: is there a large natural break in the sizes of the components, according to the scree plot? Between which two components does it occur?

There is a large break between the first and second component.

We can get some additional information about how well the PCA represents the teeth data by getting a summary of the pca object:

summary(teeth.pca)

## Importance of components:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7  
## Standard deviation 1.877 1.114 1.007 0.868 0.5577 0.3416 0.20407  
## Proportion of Variance 0.503 0.177 0.145 0.108 0.0444 0.0167 0.00595  
## Cumulative Proportion 0.503 0.680 0.825 0.933 0.9774 0.9941 1.00000

#### Question: how much variation in tooth morphology would we explain if we only retained the first component?

We would explain 50.3% of the variation in the tooth data if we only retained the first component.

#### Question: how many components would we need to retain to explain 80% or more of the variation in tooth morphology?

We would need to retain the first three compenents to explain over 80%.

### Interpreting and naming axes - PCA axes as latent variables

Calculate loadings by correlating PCA scores with the raw variables:

cor(teeth[variable.names], teeth.pca$scores) -> teeth.loadings  
teeth.loadings

## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7  
## weight 0.926 0.1216 0.2035 0.0426 0.0757 0.2823 0.00512  
## length 0.947 -0.0547 -0.0876 0.0186 0.2504 -0.1122 -0.13316  
## width.crown 0.566 0.3465 -0.2267 -0.6806 -0.2116 -0.0193 -0.00379  
## width.root 0.125 -0.4129 0.8530 -0.2869 -0.0260 -0.0584 0.00191  
## depth 0.593 -0.6473 -0.2042 0.2317 -0.3649 0.0116 -0.02932  
## surface.len 0.958 -0.1142 -0.1146 0.0733 0.1267 -0.1121 0.15133  
## thickness 0.407 0.7069 0.3622 0.3841 -0.2188 -0.0893 -0.01060

#### Question: which variables are most strongly correlated with Comp.1?

The greatest loadings are for suface length (0.958), length (0.947), and weight (0.926).

#### Question: if we were to only retain Comp.1 for further analysis, which two variables would be least well represented?

Component 1 is least correlated with width of the root and thickness.

#### Question: what should we name the latent variable represented by the first component? What about the second component? Why?

The first component is positively correlated with every variable, so it indicates size of the tooth. The second component is positively related to thickness and negatively related to depth, so it is measuring shallowness of the tooth.

Add the scores as columns in the teeth data set:

data.frame(teeth, teeth.pca$scores) -> teeth

#### Question: based no the sorted Comp.1 variable, which tooth is smallest? Which is biggest?

Tooth number 18 is smallest, and 63 is the largest.

Now import the new teeth:

data.frame(read\_excel("teeth\_fixed.xlsx","new\_teeth")) -> new.teeth

If we have the fitted PCA object we can use it to score the new teeth using predict():

predict(teeth.pca, new.teeth)

## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7  
## [1,] -1.3415 0.350 0.613 0.9646 -0.20492 0.3801 -0.2949  
## [2,] -0.0108 0.436 -1.448 -1.2315 0.45090 0.3117 0.0130  
## [3,] 0.1611 -0.307 -0.171 0.6132 -0.00186 0.0372 0.0470  
## [4,] 1.0558 0.951 -0.859 0.0717 0.15920 -0.2331 -0.1045  
## [5,] 0.3996 0.165 0.851 -0.8681 -0.24040 0.2391 -0.0672

To score the new teeth onto our PCA axes without using predict() we need the means and standard deviations for the original data set - make a vector for each:

orig.teeth.means <- c(8.11,54.10,18.09,6.08,1.51, 58.89, 10.50)  
  
orig.teeth.sds <- c(1.550,4.436,2.132,1.518,1.325,5.388,0.637)

Scale the new teeth using the means and standard deviations from the original teeth:

scale(new.teeth[variable.names], center = orig.teeth.means, scale = orig.teeth.sds) -> new.teeth.scaled  
  
new.teeth.scaled

## weight length width.crown width.root depth surface.len thickness  
## [1,] -0.17419 -0.7462 -1.121 -0.0791 -0.302 -1.0932 0.5651  
## [2,] 0.00645 0.1646 1.229 -1.0540 -0.574 0.0204 -1.0518  
## [3,] 0.07097 0.0812 -0.488 -0.2306 0.423 0.2060 0.0314  
## [4,] 0.27742 0.7755 0.704 -1.0013 -0.121 0.5772 0.5495  
## [5,] 0.50323 -0.0406 0.736 0.9354 -0.196 -0.1652 0.1413  
## attr(,"scaled:center")  
## [1] 8.11 54.10 18.09 6.08 1.51 58.89 10.50  
## attr(,"scaled:scale")  
## [1] 1.550 4.436 2.132 1.518 1.325 5.388 0.637

Enter the PCA coefficients into a matrix:

comp1.coef <- c(0.4931,0.5044,0.3015,0.0664,0.3159,0.5102,0.2169)  
  
comp2.coef <- c(0.1092,-0.0491,0.3112,-0.3708,-0.5813,-0.1026,0.6348)  
  
coeff <- cbind(comp1.coef, comp2.coef)  
  
coeff

## comp1.coef comp2.coef  
## [1,] 0.4931 0.1092  
## [2,] 0.5044 -0.0491  
## [3,] 0.3015 0.3112  
## [4,] 0.0664 -0.3708  
## [5,] 0.3159 -0.5813  
## [6,] 0.5102 -0.1026  
## [7,] 0.2169 0.6348

Now score the new teeth using the coefficients for the PCA:

data.frame(new.teeth.scaled %\*% coeff) -> new.teeth.scores  
new.teeth.scores

## comp1.coef comp2.coef  
## 1 -1.3360 0.344  
## 2 -0.0122 0.430  
## 3 0.1590 -0.309  
## 4 1.0491 0.942  
## 5 0.3962 0.160

#### Question: which tooth is the biggest according to Comp.1? Which is the smallest?

#### Question: which tooth is the deepest, according to Comp.2? Which is shallowest?

### Understanding the structure in our tooth data

Make a biplot of the PCA:

biplot(teeth.pca, pc.biplot = T)  
points(new.teeth.scores, pch = 21, bg = "red")



#### Question: based on the biplot, where on the graph would you find the largest teeth? Which tooth seems to be the largest?

The largest teeth would be on the right side of the graph, and tooth 63 appears to be the largest.

Assign the scores for Comp.1 and Comp.2 to the teeth data, so that you can open the data set and sort on each.

data.frame(teeth, teeth.pca$scores[,1:2]) -> teeth

#### Question: describe what tooth 84 should look like, based on the biplot. Should it be large, medium, or small? Should it be thick or thin? Should it be strongly curved (big depth) or shallow? Should it have a wide root or a narrow one? Then look at tooth 84 in the data set and see if it has the measurements you expected.

Tooth 84 is medium sized, but it should be thin, strongly curved and have a wide root.

#### Question: what multi-way pattern of correlation is Comp.1 illustrating? That is, which variables are all inter-correlated, according to Comp.1?

#### Question: same question for Comp.2 - which pattern of inter-correlation does Comp.2 illustrate?

### The implication of reducing dimensionality

Calculate the explained variance for each variable by squaring the loadings:

teeth.loadings^2 -> teeth.var.explained  
teeth.var.explained

## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7  
## weight 0.8566 0.01478 0.04140 0.001815 0.005724 0.079669 2.63e-05  
## length 0.8960 0.00299 0.00767 0.000344 0.062689 0.012590 1.77e-02  
## width.crown 0.3202 0.12006 0.05140 0.463160 0.044769 0.000374 1.44e-05  
## width.root 0.0155 0.17050 0.72757 0.082335 0.000675 0.003405 3.65e-06  
## depth 0.3515 0.41900 0.04168 0.053685 0.133175 0.000134 8.60e-04  
## surface.len 0.9169 0.01305 0.01314 0.005367 0.016063 0.012556 2.29e-02  
## thickness 0.1657 0.49964 0.13116 0.147549 0.047883 0.007982 1.12e-04

#### Question: which variable is best explained by Comp.1? How much of its variation is explained by Comp.1?

Since these are squared correlations, the same variables that had high loadings are best explained. The highest is length, with 86.1% of its variance explained by Comp.1

Calculate the communalities (i.e. running sums of variance explained across the PC axes). First make the ones matrix, and then multiply the explained variation values by it to get communalities:

round(upper.tri(matrix(0,7,7), diag = T)) -> ones.matrix  
  
teeth.var.explained %\*% ones.matrix -> teeth.communalities  
  
colnames(teeth.communalities) <- paste("First", seq(1,7), sep = ".")  
  
teeth.communalities

## First.1 First.2 First.3 First.4 First.5 First.6 First.7  
## weight 0.8566 0.871 0.913 0.915 0.920 1.000 1  
## length 0.8960 0.899 0.907 0.907 0.970 0.982 1  
## width.crown 0.3202 0.440 0.492 0.955 1.000 1.000 1  
## width.root 0.0155 0.186 0.914 0.996 0.997 1.000 1  
## depth 0.3515 0.770 0.812 0.866 0.999 0.999 1  
## surface.len 0.9169 0.930 0.943 0.948 0.965 0.977 1  
## thickness 0.1657 0.665 0.796 0.944 0.992 1.000 1

#### Question: how many components need to be retained to encompass 80% of the variation in length? How many are needed to encompass 80% or more of the variation in width.root?

Component 1 explains more than 80% of the variation in length by itself. Root width would require three components to represent more than 80% of its variation.

### Design considerations

#### Question: did we have enough data to conduct this PCA? Refer to the sample size guidelines from the book (by way of lecture).

We did, because we have several loadings that were over 0.6.

Check that the PC axes are independent by calculating the correlations between their scores:

cor(teeth.pca$scores)

## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7  
## Comp.1 1.00e+00 -1.70e-16 -2.06e-16 -8.00e-19 -1.11e-16 -2.11e-15 -1.31e-16  
## Comp.2 -1.70e-16 1.00e+00 2.60e-15 8.60e-16 -5.60e-16 -3.26e-16 8.50e-16  
## Comp.3 -2.06e-16 2.60e-15 1.00e+00 -5.27e-16 2.77e-16 6.13e-16 3.66e-16  
## Comp.4 -8.00e-19 8.60e-16 -5.27e-16 1.00e+00 1.01e-15 3.47e-16 -2.20e-15  
## Comp.5 -1.11e-16 -5.60e-16 2.77e-16 1.01e-15 1.00e+00 5.27e-16 2.70e-15  
## Comp.6 -2.11e-15 -3.26e-16 6.13e-16 3.47e-16 5.27e-16 1.00e+00 1.57e-15  
## Comp.7 -1.31e-16 8.50e-16 3.66e-16 -2.20e-15 2.70e-15 1.57e-15 1.00e+00

Be sure to look at the large printout of the biplot, with teeth placed along PC 1 and PC 2, so that you can see what each axis is telling you about the patterns of variation in the teeth.

#### Question: after looking at how the teeth are distributed across the PCA axes, do you see any characteristics of the teeth that we didn’t represent well with our variables? Name one or two additional measurements we could have made to account for anything you think our PCA missed.