NCSU ST 503 Discussion 10

Probem 11.6 Faraway, Julian J. Linear Models with R CRC Press.

Bruce Campbell

10.6 PCA analysis of kanga dataset

The dataset kanga contains data on the skulls of historical kangaroo specimens.

(a) Compute a PCA on (the 18 skull measurements. You will need to exclude observations with missing values. What percentage of variation is explained by the first principal component?

```
## Importance of components:
##                             PC1      PC2      PC3      PC4      PC5
## Standard deviation     288.0382 69.51124 30.74720 27.85580 21.73015
## Proportion of Variance   0.9003  0.05243  0.01026  0.00842  0.00512
## Cumulative Proportion    0.9003  0.95269  0.96295  0.97136  0.97649
##                             PC6      PC7     PC8      PC9     PC10
## Standard deviation     19.42356 17.28247 16.6247 14.52310 13.98826
## Proportion of Variance  0.00409  0.00324  0.0030  0.00229  0.00212
## Cumulative Proportion   0.98058  0.98382  0.9868  0.98911  0.99123
##                            PC11     PC12     PC13     PC14    PC15    PC16
## Standard deviation     12.35253 12.07402 11.94245 10.82939 10.0735 8.46081
## Proportion of Variance  0.00166  0.00158  0.00155  0.00127  0.0011 0.00078
## Cumulative Proportion   0.99289  0.99447  0.99602  0.99729  0.9984 0.99917
##                           PC17    PC18
## Standard deviation     7.16825 5.01246
## Proportion of Variance 0.00056 0.00027
## Cumulative Proportion  0.99973 1.00000
```

47 data elements were removed due to missing values in the measurement dimensions. We see that %90 of variance in the measurements is explained by the first principal component.

(b) Provide the loadings for the first principal component. What variables are prominent?

The loadings for a principal component $\mathbf{u}_i$ are the values of the dimensions u_{ij} , in our case the measurements. We note that the textbook uses r method prcomp to perform principal components and that it gets loadings from the rot matrix. There is another r function in common use for pca - princomp. This method has a loading structure in the output. We tested this method and got a vector similar to the other method except all the signs were reversed. This is OK because the direction is the same - i.e. if we project all the data points on the first version, we'll get the same points as if we had projected on the second version. We also note there is some confusion on the difference between eigenvectors and loadings. The rot matrix is orthogonal - we checked this for a few values

```
t(pca.kanga$rotation[,1]) %*%  pca.kanga$rotation[,1]
t(pca.kanga$rotation[,3]) %*%  pca.kanga$rotation[,4]
t(pca.kanga$rotation[,1]) %*%  pca.kanga$rotation[,2]
```

Table 1: First Principal Component

	first.pc.loadings
basilar.length	0.484
occipitonasal.length	0.456
palate.length	0.366
palate.width	0.084
nasal.length	0.248
nasal.width	0.075
squamosal.depth	0.064
lacrymal.width	0.119
zygomatic.width	0.207
orbital.width	0.014
.rostral.width	0.106
occipital.depth	0.178
crest.width	-0.082
foramina.length	0.01
mandible.length	0.436
mandible.width	0.03
mandible.depth	0.058
ramus.height	0.209

We note that the following measurements all have loadings greater than .2

$$\{basilar.length,\ occipitonasal.length,\ palate.length,\ nasal.length,\ mandible.length,\ zygomatic.widt$$

(c) Repeat the PCA but with the variables all scaled to the same standard deviation. How do the percentage of variation explained and the first principal component differ from those found in the previous PCA?

PCA of scaled measurements

```
## Importance of components:
##                           PC1     PC2    PC3    PC4    PC5     PC6     PC7
## Standard deviation     3.5321 1.30672 1.1006 0.8443 0.6463 0.56426 0.51064
## Proportion of Variance 0.6931 0.09486 0.0673 0.0396 0.0232 0.01769 0.01449
## Cumulative Proportion  0.6931 0.78796 0.8553 0.8949 0.9181 0.93575 0.95024
##                            PC8     PC9   PC10    PC11   PC12    PC13
## Standard deviation     0.45185 0.43863 0.3723 0.30491 0.2815 0.24345
## Proportion of Variance 0.01134 0.01069 0.0077 0.00517 0.0044 0.00329
## Cumulative Proportion  0.96158 0.97227 0.9800 0.98514 0.9895 0.99283
##                           PC14    PC15    PC16    PC17    PC18
## Standard deviation     0.22317 0.18583 0.15031 0.11849 0.08949
## Proportion of Variance 0.00277 0.00192 0.00126 0.00078 0.00044
## Cumulative Proportion  0.99560 0.99752 0.99878 0.99956 1.00000
```

After scaling the proportion of variance explained by the first principal component has dropped to .69

(d) Give an interpretation of the second principal component.

Table 2: Second Principal Component

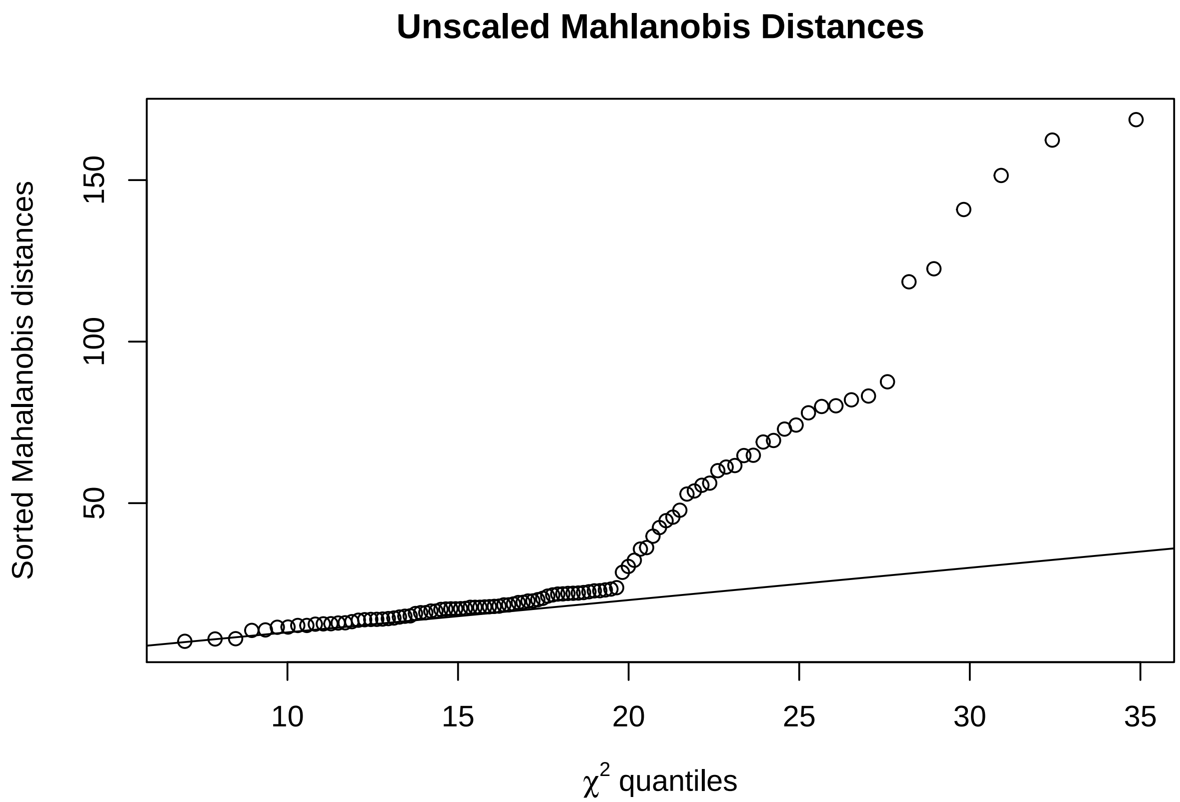
	first.pc.loadings
basilar.length	-0.138
occipitonasal.length	0.414
palate.length	-0.002
palate.width	-0.023
nasal.length	0.584
nasal.width	0.127
squamosal.depth	-0.105
lacrymal.width	-0.04
zygomatic.width	-0.41
orbital.width	0.001
.rostral.width	-0.063
occipital.depth	-0.079
crest.width	-0.245
foramina.length	0.061
mandible.length	-0.212
mandible.width	-0.092
mandible.depth	-0.106
ramus.height	-0.365

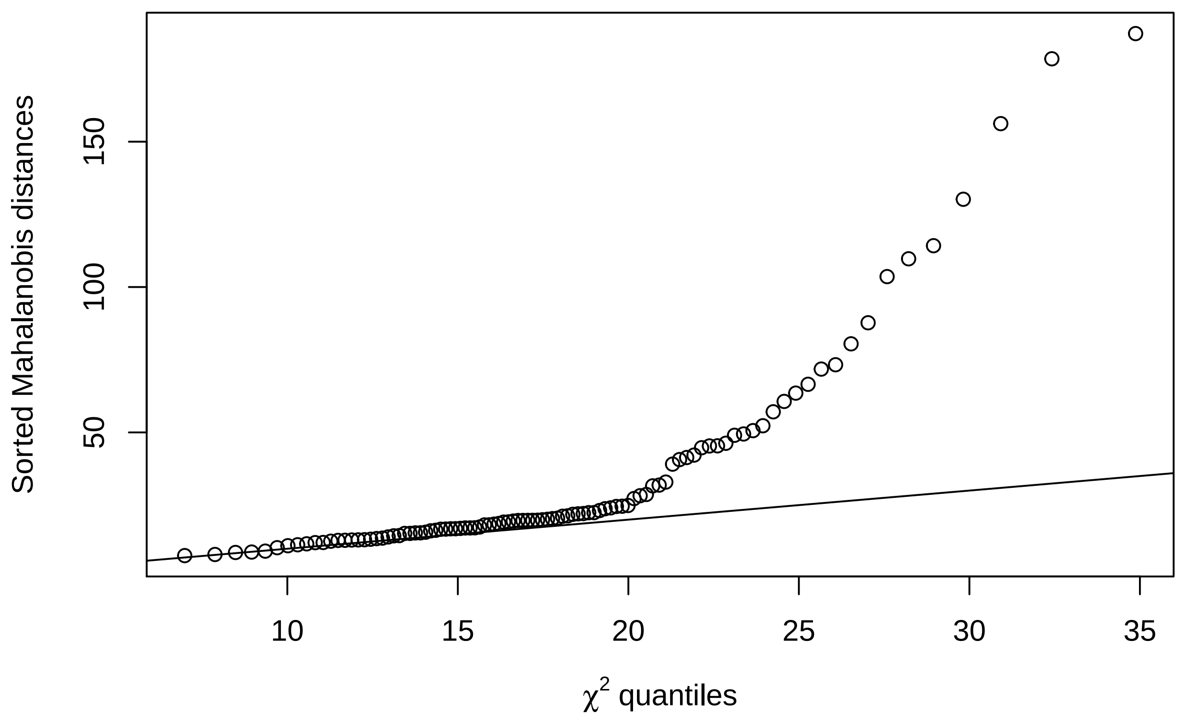
As evidenced by the loadings, the first principal component mainly account for variation in length feature. The second principal component is primarily a contrast between the variables

$\{occipitonasal.length,\ \ nasal.length\}$ and $\{ramus.height,\ \ crest.width\}$

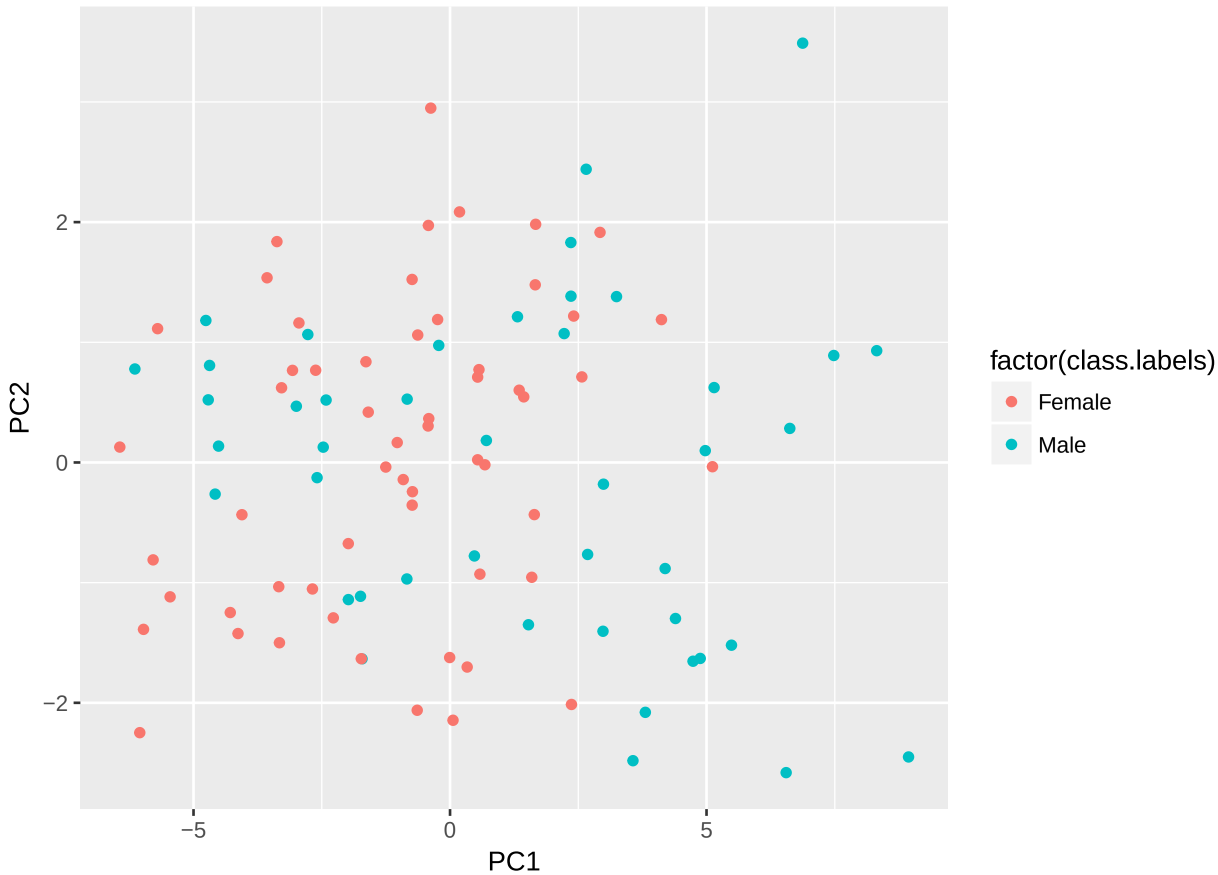
(e) Compute the Mahalanobis distances and plot appropriately to check for outliers.

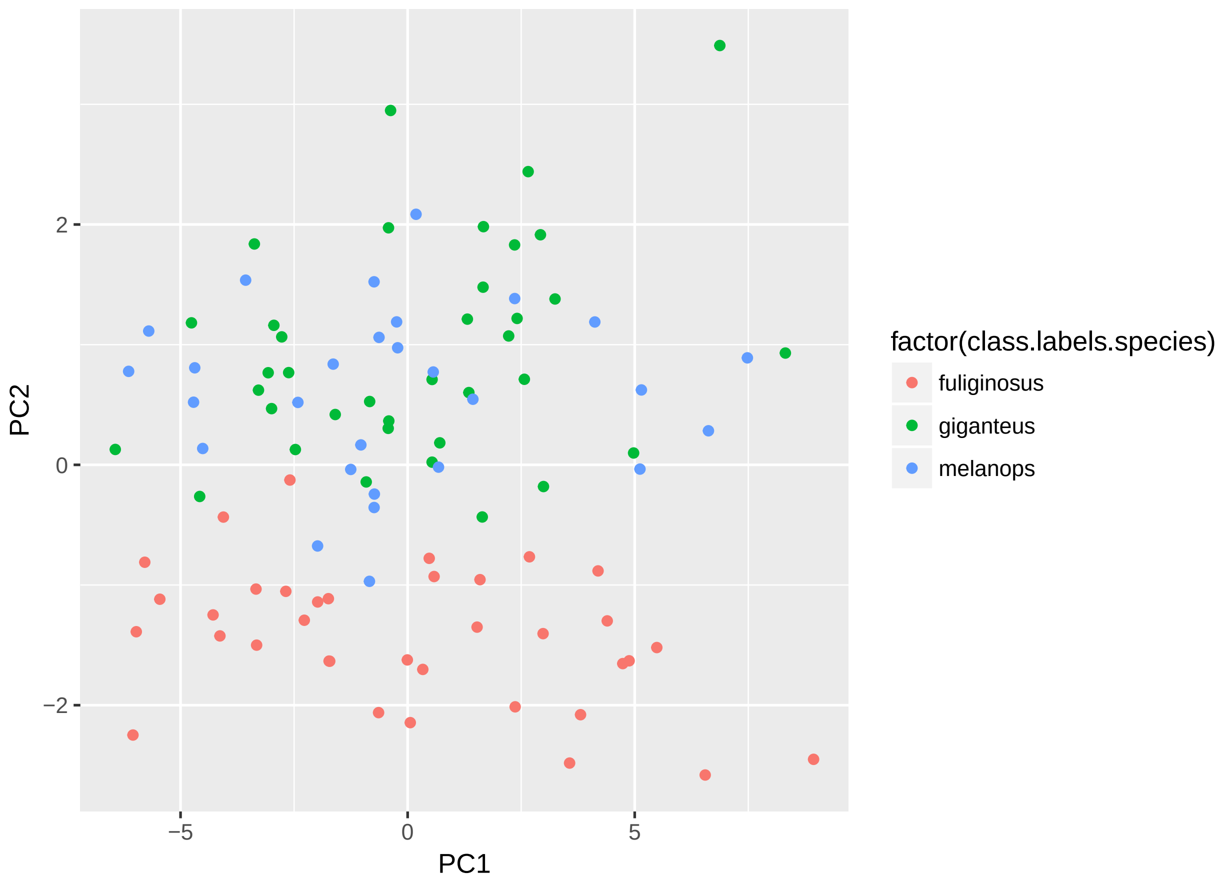
We calculate the distances in the unscaled and scaled data




Sorted Mahalanobis distances
50
100
150
10
15
20
25
30
35
χ^2 quantiles

(f) Make a scatterplot of the first and second principal components using a different plotting symbol depending on the sex of the specimen. Do you think these two components would be effective in determining the sex of a skull?





We see that the measurements do not allow for a linear classifier for discriminating sex via the first two PCA projections. We can discriminate the species fulignosus from giganteus and melanops.