STA 663: Homework #4

Brian N. White

3/3/2020

```
#import the data
wdbc_F <- read.csv("wdbc_F.csv")
```

Question 1

As described in the paper in question, one way in which breast cancer is diagnosed is through a biopsy. A less invasive alternative to this procedure is a fine needle aspiration (FNA). FNAs enable one to examine a small sample of tumor tissue. This tissue is then examined by physicians who make a subjective classification of the sample as cancerous or not. The motivation of the experiment in question was to develop a more objective means of detecting the presence of breast tumors via fine needle aspirations (FNA).

Problem 2

Each observation is an image. In particular, each image is of a stained drop of fluid viewed under microscope taken from the FNA of a breast cancer tumor. The sample size is 568.

Problem 3

There are 32 variables. Of these variables, 30 are engineered from the ten variables (i.e. not present in the data set) radius, perimeter, area, compactness, smoothness, concavity, concave points, symmetry, fractal dimension and texture. These ten refer to features of the nuclei present in each image. In particular, the mean, max (i.e. ‘worst’) and standard deviation of these ten are calculated for each image which gives the aforementioned 30 variables. The remaining 2 variables are the observation ID, ‘id’, and the binary categorical variable, ‘diagnosis’.

Problem 4

The variable ‘radius’ (not present in the data set) refers to the radius of an individual nucleus in an image. As described in the paper in question, this value is calculated by taking the mean of the length of the radial line segments defined by the centroid of the snake. Each image will have a ‘radius’ value for each nucleus. Thus, ‘radius_mean’, ‘radius_se’ and ‘radius_worst’ give the mean, standard deviation, and max, respectively, of the radius values for each image.

Problem 5

As mentioned above, the procedue used to obtain the cells from each image is called a fine needle aspiration (FNA). A FNA is performed by first extracting a drop of fluid from a breast tumor using a fine needle. This drop is then placed onto a glass slide and stained. Then the slide is examined under microscope and and photographed via specialized software.

Problem 6

```
pca_tumor=prcomp(wdbc_F[,3:32], scale=TRUE, center=TRUE)
```

Problem 7

The scree plot is displayed below. A threshold of .9 was used and let me to select the first 7 principle components. Note that this can be confirmed visually in the sree plot.

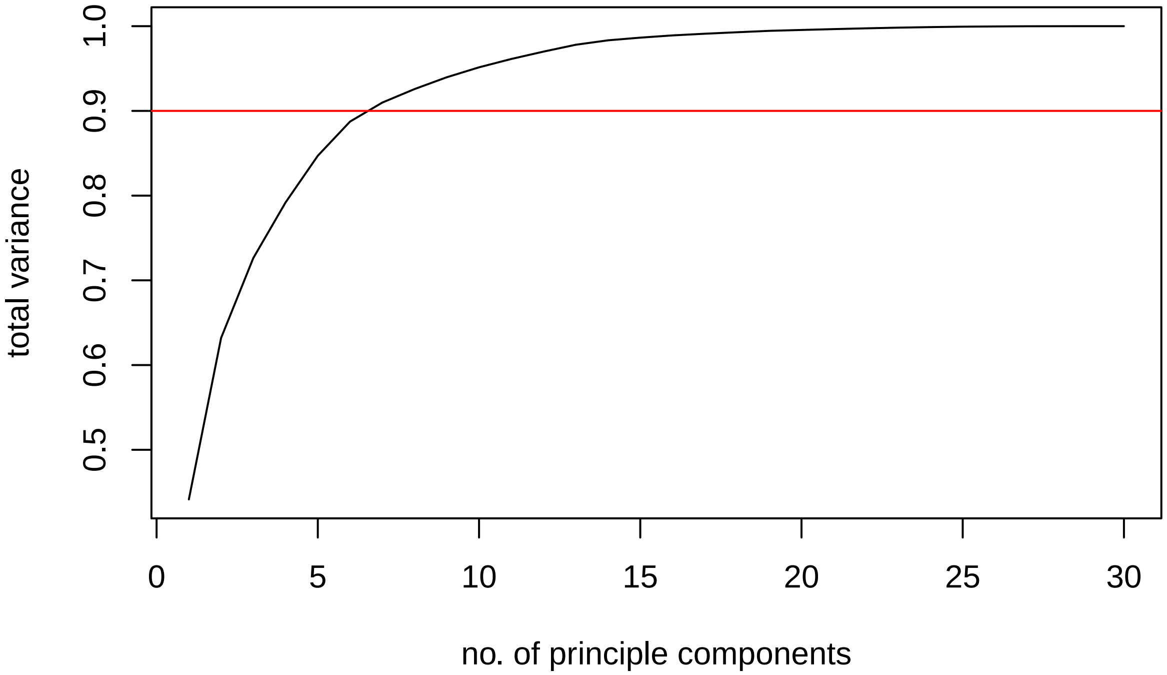
```
cum_var=summary(pca_tumor)$importance[3, ]
cum_var
```

```
##     PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10
## 0.44141 0.63197 0.72638 0.79195 0.84707 0.88744 0.90982 0.92576 0.93972 0.95143
##    PC11    PC12    PC13    PC14    PC15    PC16    PC17    PC18    PC19    PC20
## 0.96129 0.96999 0.97809 0.98332 0.98646 0.98911 0.99110 0.99284 0.99451 0.99555
##    PC21    PC22    PC23    PC24    PC25    PC26    PC27    PC28    PC29    PC30
## 0.99655 0.99747 0.99829 0.99890 0.99942 0.99969 0.99992 0.99997 1.00000 1.00000
```

```
P=1:length(cum_var) #the number of principle components

plot(cum_var~P, type="l", xlab="no. of principle components", ylab="total variance")

abline(.90, 0, col="red") #cumulative variance threshold of .9
```



```
which(cum_var>=.9)
```

```
##  PC7  PC8  PC9 PC10 PC11 PC12 PC13 PC14 PC15 PC16 PC17 PC18 PC19 PC20 PC21 PC22
##    7    8    9   10   11   12   13   14   15   16   17   18   19   20   21   22
## PC23 PC24 PC25 PC26 PC27 PC28 PC29 PC30
##   23   24   25   26   27   28   29   30
```

Problem 8

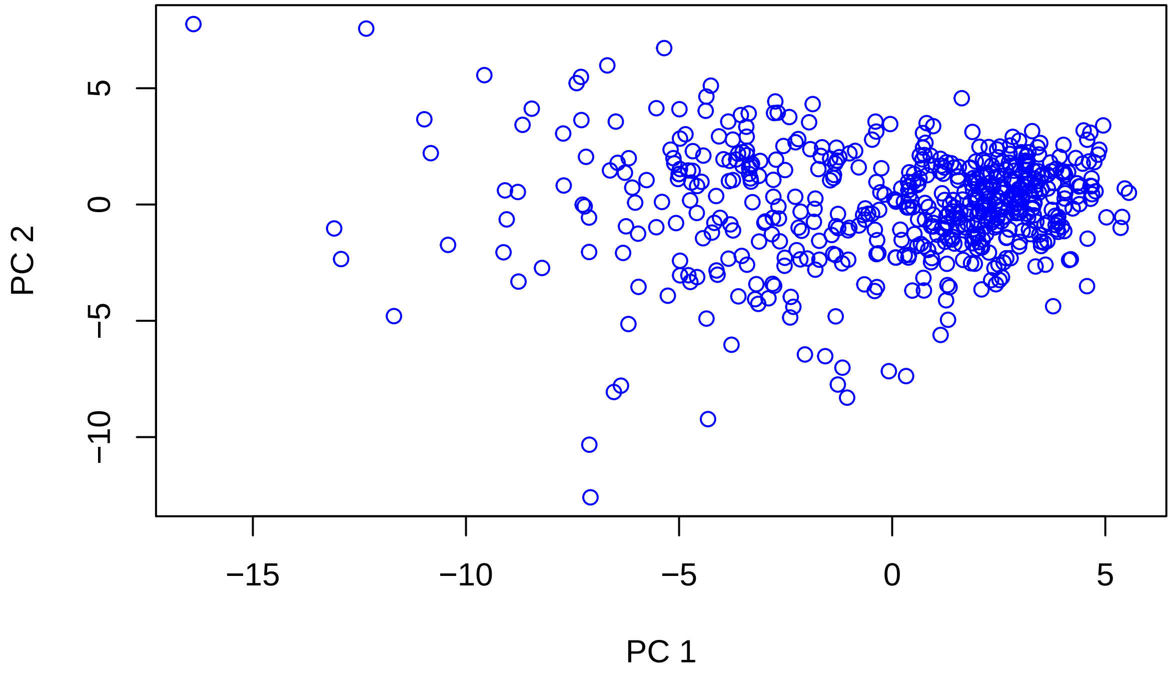
The new data set, created from the first 7 principle components is computed and printed below as 'Y_reduced'.

```
r=pca_tumor$rotation
wdbc_F.c=scale(wdbc_F[, 3:32], center=T, scale=T)
Y_reduced=wdbc_F.c%*%r[,1:7]
head(Y_reduced)
```

```
##              PC1        PC2        PC3       PC4         PC5         PC6
## [1,] -2.414217   3.760654 -0.5323563 1.1464043 -0.59203853  0.05100512
## [2,] -5.763319   1.050563 -0.5571854 0.9486727  0.21344557  0.57792234
## [3,] -7.105538 -10.326125 -3.2240362 0.1497268  2.98827366  3.07283497
## [4,] -3.957647   1.944834  1.4086806 2.9821889 -0.47876096 -1.17095964
## [5,] -2.379174  -3.972076 -2.9219768 0.9823938  1.09169909 -0.43347591
## [6,] -2.267395   2.675690 -1.6501372 0.1817093 -0.02731665 -0.11703601
##              PC7
## [1,]  0.03936053
## [2,] -0.62358126
## [3,]  1.53244465
## [4,] -0.91084348
## [5,]  0.51780929
## [6,] -0.29015223
```

Problem 9

```
plot(Y_reduced[, 2]~Y_reduced[, 1], col="blue", xlab="PC 1", ylab="PC 2")
```



Problem 10

As can be seen in the plot below, the first and second principle components are fairly effective in discriminating between benign and malignent tumor cells.

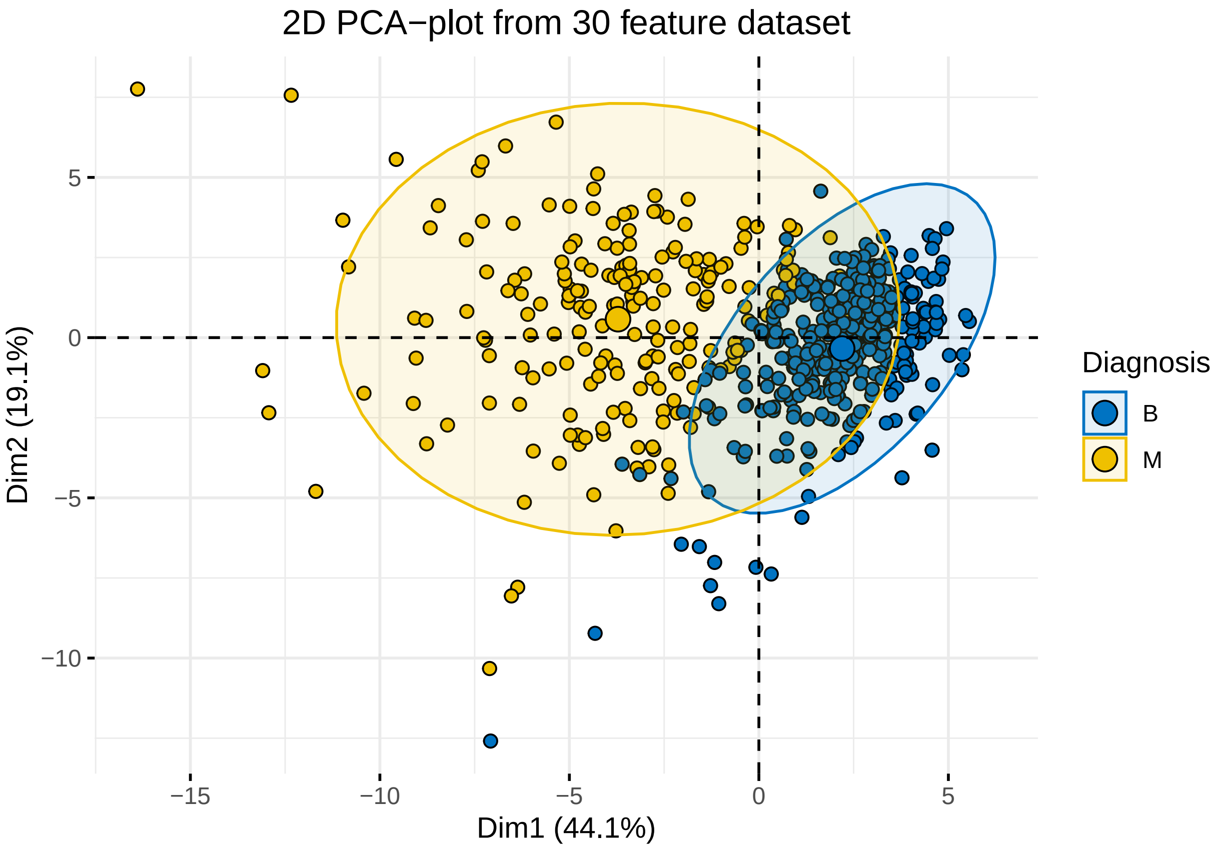
```
library("factoextra")
```

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

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
fviz_pca_ind(pca_tumor, geom.ind = "point", pointshape = 21,

pointsize = 2,
fill.ind = wdbc_F$diagnosis,
col.ind = "black",
palette = "jco",
addEllipses = TRUE,
label = "var",
col.var = "black",
repel = TRUE,
legend.title = "Diagnosis") +
ggtitle("2D PCA-plot from 30 feature dataset") +
theme(plot.title = element_text(hjust = 0.5))
```


2D PCA−plot from 30 feature dataset
Dim2 (19.1%)
Dim1 (44.1%)
Diagnosis
B
M