

MATH 308 Assignment 4

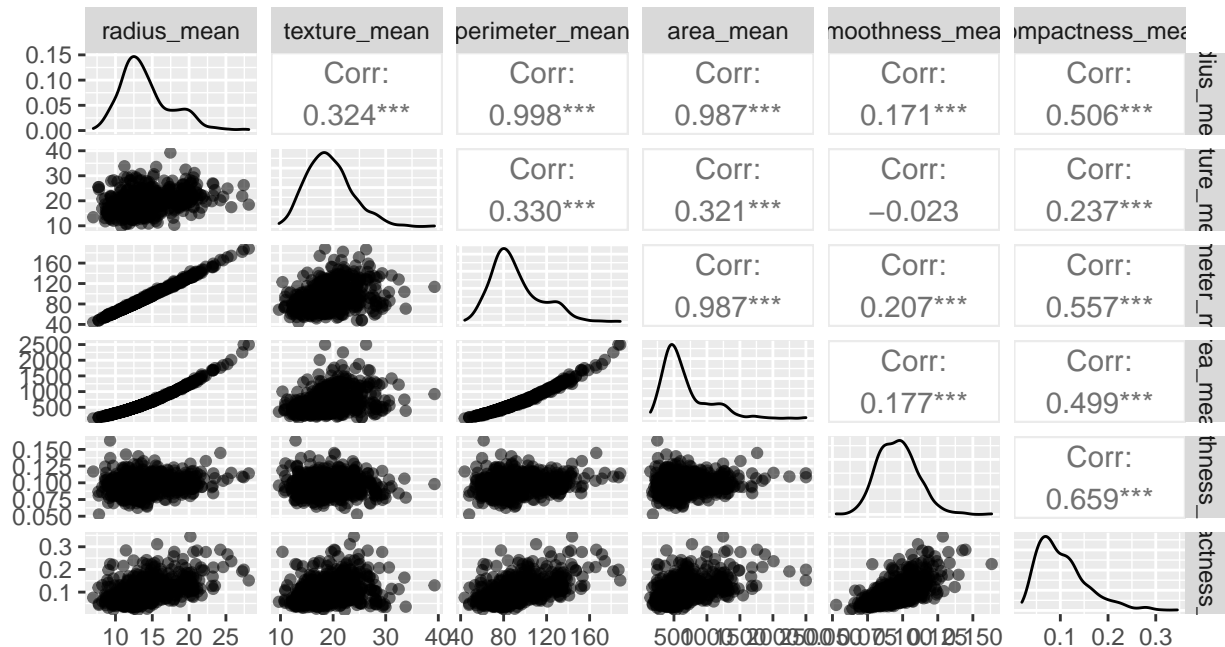
Breast cancer is the most common malignancy among women, accounting for nearly 1 in 3 cancers diagnosed among women in the United States, and it is the second leading cause of cancer death among women. Breast Cancer occurs as a result of abnormal growth of cells in the breast tissue, commonly referred to as a Tumor. The task given by the project is to perform principal component analysis (PCA) for the given continuous data that stored in the dataset Breast-Cancer-Wisconsin created by Dr. William H. Wolberg. Thus, the first thing we did is to remove all the categorical data and useless data columns, such as column 'class'. Those column contains the categorical data as 1 or 2. In the first run, we chose to use the mean for our PCA on the data.

The real-valued features (continuous data) are computed for each cell nucleus:

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness ($\text{perimeter}^2 / \text{area} - 1.0$)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension ("coastline approximation" - 1)

To better understand the correlation between variables, we present the pairwise plot. We only consider the first 6 variables here for the readability of the figure.

```
ggpairs(sub_df[3:8], aes(alpha = 0.1))
```



We have a look of mean

```
## Mean
apply(sub_df[3:12] , 2, mean)
```

```
##          radius_mean      texture_mean      perimeter_mean
##          14.12729174          19.28964851          91.96903339
##          area_mean      smoothness_mean      compactness_mean
##          654.88910369          0.09636028          0.10434098
##          concavity_mean      concave.points_mean      symmetry_mean
##          0.08879932          0.04891915          0.18116186
## fractal_dimension_mean
##          0.06279761
```

and variance as well.

```
## Variance
apply(sub_df[3:12] , 2, var)
```

```
##          radius_mean      texture_mean      perimeter_mean
##          1.241892e+01      1.849891e+01      5.904405e+02
##          area_mean      smoothness_mean      compactness_mean
##          1.238436e+05      1.977997e-04      2.789187e-03
##          concavity_mean      concave.points_mean      symmetry_mean
##          6.355248e-03      1.505661e-03      7.515428e-04
## fractal_dimension_mean
##          4.984872e-05
```

Since the variables all have different and incomparable scales, we need to scale them to unit variance when finding the principle components.

```
br_cancer_mean.pca <- prcomp(sub_df[3:12] , scale=TRUE)
summary(br_cancer_mean.pca)
```

```
## Importance of components:
##          PC1      PC2      PC3      PC4      PC5      PC6      PC7
## Standard deviation      2.3406 1.5870 0.93841 0.7064 0.61036 0.35234 0.28299
## Proportion of Variance 0.5479 0.2519 0.08806 0.0499 0.03725 0.01241 0.00801
## Cumulative Proportion 0.5479 0.7997 0.88779 0.9377 0.97495 0.98736 0.99537
##          PC8      PC9      PC10
## Standard deviation      0.18679 0.10552 0.01680
## Proportion of Variance 0.00349 0.00111 0.00003
## Cumulative Proportion 0.99886 0.99997 1.00000
```

```
br_cancer_all.pca <- prcomp(df[3:32] , scale=TRUE)
summary(br_cancer_all.pca)
```

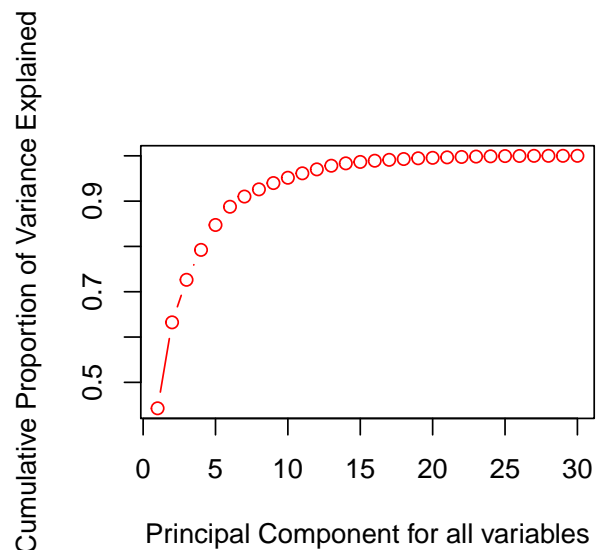
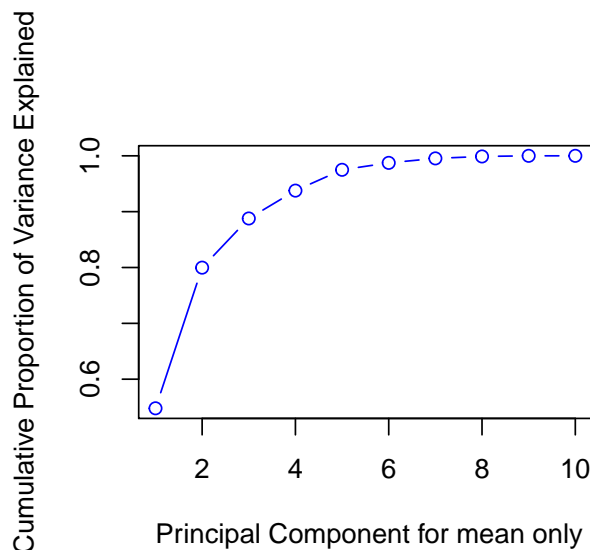
```
## Importance of components:
##          PC1      PC2      PC3      PC4      PC5      PC6      PC7
## Standard deviation      3.6444 2.3857 1.67867 1.40735 1.28403 1.09880 0.82172
## Proportion of Variance 0.4427 0.1897 0.09393 0.06602 0.05496 0.04025 0.02251
## Cumulative Proportion 0.4427 0.6324 0.72636 0.79239 0.84734 0.88759 0.91010
##          PC8      PC9      PC10      PC11      PC12      PC13      PC14
## Standard deviation      0.69037 0.6457 0.59219 0.5421 0.51104 0.49128 0.39624
## Proportion of Variance 0.01589 0.0139 0.01169 0.0098 0.00871 0.00805 0.00523
## Cumulative Proportion 0.92598 0.9399 0.95157 0.9614 0.97007 0.97812 0.98335
##          PC15      PC16      PC17      PC18      PC19      PC20      PC21
## Standard deviation      0.30681 0.28260 0.24372 0.22939 0.22244 0.17652 0.1731
## Proportion of Variance 0.00314 0.00266 0.00198 0.00175 0.00165 0.00104 0.0010
```

```
## Cumulative Proportion  0.98649 0.98915 0.99113 0.99288 0.99453 0.99557 0.9966
##                        PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation     0.16565 0.15602 0.1344 0.12442 0.09043 0.08307 0.03987
## Proportion of Variance 0.00091 0.00081 0.0006 0.00052 0.00027 0.00023 0.00005
## Cumulative Proportion  0.99749 0.99830 0.9989 0.99942 0.99969 0.99992 0.99997
##                        PC29    PC30
## Standard deviation     0.02736 0.01153
## Proportion of Variance 0.00002 0.00000
## Cumulative Proportion  1.00000 1.00000
```

We have a look of cumulative proportion of variance explained. Notice that the first two components only explain 63% of the total variance, which means more components are needed for data analysis. Applying ‘the 80% Rule’, we need to consider the first five components.

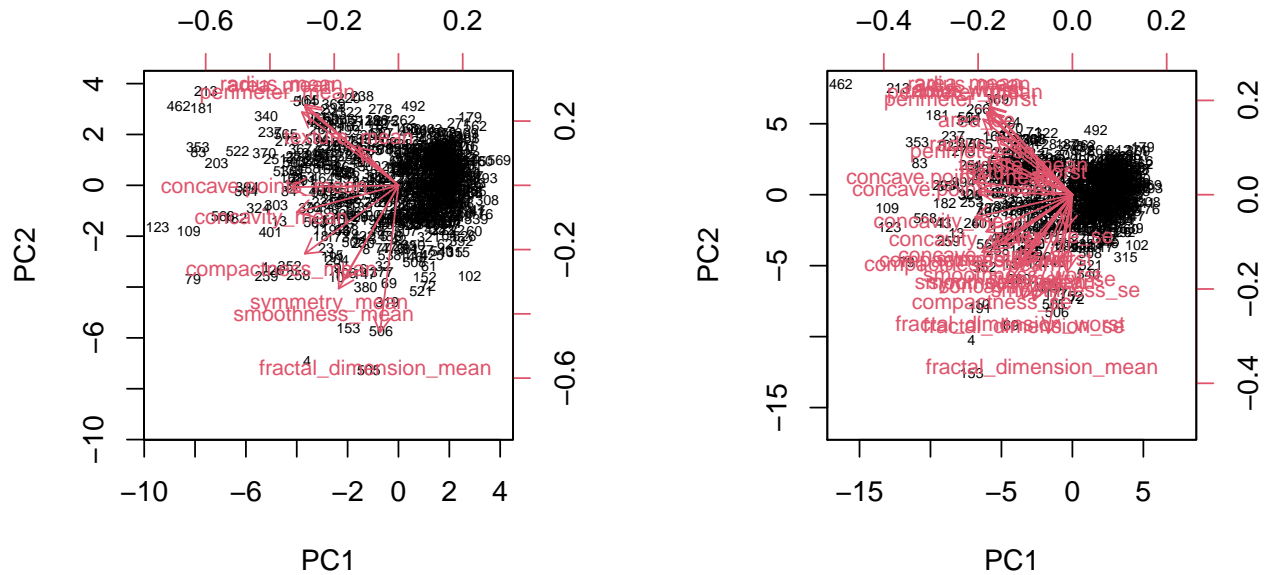
The plots below are the proportion of variance by every principal component only for mean and for all variables.

```
par(mfrow =c(1,2))
plot(cumsum(br_cancer_mean.pca$sdev^2/(sum(br_cancer_mean.pca$sdev^2))),
     xlab=" Principal Component for mean only",
     ylab ="Cumulative Proportion of Variance Explained",type='b',col =" blue")
plot(cumsum(br_cancer_all.pca$sdev^2/(sum(br_cancer_all.pca$sdev^2))),
     xlab="Principal Component for all variables",
     ylab="Cumulative Proportion of Variance Explained", type='b',col =" red")
```



Check the biplot

```
par(mfrow =c(1,2))
biplot(br_cancer_mean.pca,scale=0, cex = c(0.5, 0.75))
biplot(br_cancer_all.pca,scale=0, cex = c(0.5, 0.75))
```

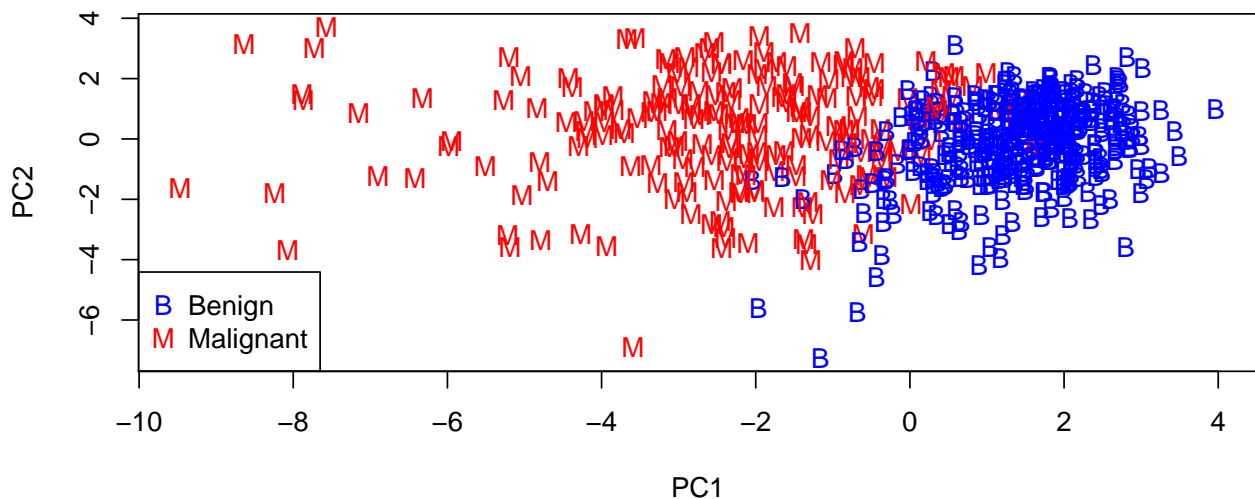


We found a regular pattern that for the first main Principle Component, if we only take consideration in mean, PC is positive. In that way, considering all Principle Component of 30 variables, they maintain the corresponding positive or negative signs in most cases. In addition, standard deviations and worsts are strongly correlated with mean. That's the reason why we chose mean to summarize our dataset. By checking the biplots, for the first plot (using mean), PC1 is associated with concave point_mean, and PC2 is associated with radius_mean. The second plots show the same association with PC1 and PC2. This result also indicates that mean value is sufficient to represent the data.

Next, we plot the first two projected X . In the first graph, the X is from PCA at which only mean is considered. In the second graph, mean, standard deviation and "worst" of all 10 features is considered. We find 10 mean variables are as good as 30 all mean, sd, worst variables in providing a separation.

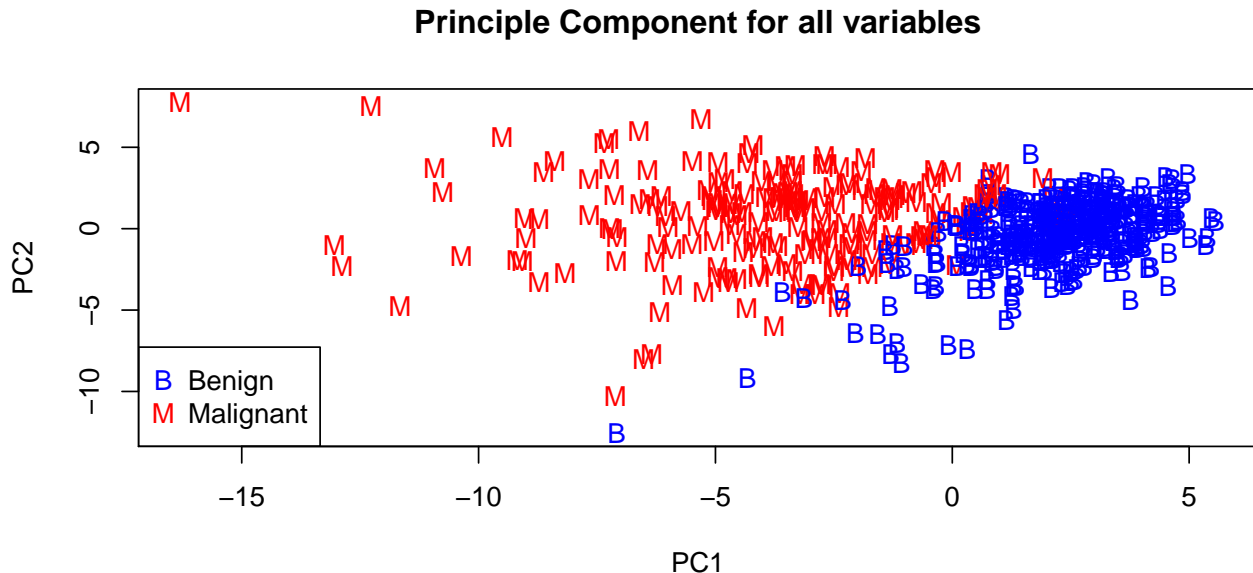
```
plot(br_cancer_mean.pca$x[, 1:2], pch = ifelse(sub_df[, "diagnosis"] == "B", "B", "M"),
     col = ifelse(sub_df[, "diagnosis"] == "B", "blue", "red"),
     main="Principle Component for mean only")
legend("bottomleft", legend = c("Benign", "Malignant"),
     col = c("blue", "red"), pch = c("B", "M"))
```

Principle Component for mean only



In this plot, we have reduced the dimensions to 2, which keeps less information. Observing the plot above, we can see that these two variables can be separated by a line with Benign on the right side and Malignant on the other side. If we use the first five components, there will be much more variables and dimensions to analyze. However, using two dimensions can help us summarize the information better with lots of data existing.

```
plot(br_cancer_all.pca$x[, 1:2], pch = ifelse(sub_df[, "diagnosis"] == "B", "B", "M"),
     col = ifelse(sub_df[, "diagnosis"] == "B", "blue", "red"),
     main="Principle Component for all variables")
legend("bottomleft", legend = c("Benign", "Malignant"),
     col = c("blue", "red"), pch = c("B", "M"))
```



Notice that in both cases, there exists a straight line that separates the benign and malignant observations. However, we see more overlapping in the case with only means. This suggests the prediction based on means may not be as reliable as when all variables are considered.

We found that PCA is a powerful technique to reduce the dimensions of data. In our specific case, the dimension id reduced from 32 to 2, without losing significant interpretability and information. Two principle components can explain about 80% variance present in the data. When we plot the data in the first two components, an obvious boundary is found to separate the two types of cancers. This suggests that a linear classification model can be established based on this analysis.