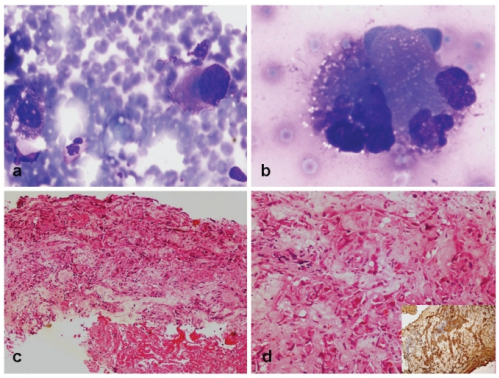
Breast Cancer Wisconsin (Diagnostic)

L. Favero

Contents

1 Information about the dataset

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29 "Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image."



Ten real-valued features are computed for each cell nucleus:

Name of the variables	type	Description
1) 'radius'	num	distances from center to points on the perimeter
2) 'texture'	num	standard deviation of gray-scale values
3) 'perimeter'	num	perimeter of the nucleus
4) 'area'	num	area of the nucleus
5) 'smoothness'	num	local variation in radius lengths
6) 'compactness'	num	$perimeter^2/area - 1.0$
7) 'concavity'	num	severity of concave portions of the contour
8) 'concave.points'	num	number of concave portions of the contour
9) 'symmetry'	num	symmetry of the nucleus
10)'fractal_dimension'	num	$coastlineapproximation - 1$

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. All feature values are recorded with four significant digits.

The aim is to **predict whether the cancer is benign or malignant**

2 Preprocessing

2.1 Load

Load libraries

```
library(GGally)       # for ggpairs
```

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

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(ggfortify)    # for autoplot
library(ggplot2)      # for ggplot
library('MASS')       # for the glm model selection
```

Load data

```
df<-read.csv('/Users/lucile/Library/Mobile Documents/com~apple~CloudDocs/STUDY
```

2.2 First look into the data

```
str(df)
```

```
## 'data.frame':    569 obs. of  33 variables:
##  $ id                     : int  842302 842517 84300903 84348301 84358402 8
##  $ diagnosis              : Factor w/ 2 levels "B","M": 2 2 2 2 2 2 2 2 2 2
##  $ radius_mean            : num  18 20.6 19.7 11.4 20.3 ...
##  $ texture_mean           : num  10.4 17.8 21.2 20.4 14.3 ...
##  $ perimeter_mean         : num  122.8 132.9 130 77.6 135.1 ...
##  $ area_mean              : num  1001 1326 1203 386 1297 ...
##  $ smoothness_mean        : num  0.1184 0.0847 0.1096 0.1425 0.1003 ...
##  $ compactness_mean       : num  0.2776 0.0786 0.1599 0.2839 0.1328 ...
##  $ concavity_mean         : num  0.3001 0.0869 0.1974 0.2414 0.198 ...
##  $ concave.points_mean    : num  0.1471 0.0702 0.1279 0.1052 0.1043 ...
##  $ symmetry_mean          : num  0.242 0.181 0.207 0.26 0.181 ...
##  $ fractal_dimension_mean : num  0.0787 0.0567 0.06 0.0974 0.0588 ...
##  $ radius_se              : num  1.095 0.543 0.746 0.496 0.757 ...
##  $ texture_se             : num  0.905 0.734 0.787 1.156 0.781 ...
##  $ perimeter_se           : num  8.59 3.4 4.58 3.44 5.44 ...
##  $ area_se                : num  153.4 74.1 94 27.2 94.4 ...
##  $ smoothness_se          : num  0.0064 0.00522 0.00615 0.00911 0.01149 ..
##  $ compactness_se         : num  0.049 0.0131 0.0401 0.0746 0.0246 ...
##  $ concavity_se           : num  0.0537 0.0186 0.0383 0.0566 0.0569 ...
##  $ concave.points_se      : num  0.0159 0.0134 0.0206 0.0187 0.0188 ...
##  $ symmetry_se            : num  0.03 0.0139 0.0225 0.0596 0.0176 ...
##  $ fractal_dimension_se   : num  0.00619 0.00353 0.00457 0.00921 0.00511 .
##  $ radius_worst           : num  25.4 25 23.6 14.9 22.5 ...
##  $ texture_worst          : num  17.3 23.4 25.5 26.5 16.7 ...
##  $ perimeter_worst        : num  184.6 158.8 152.5 98.9 152.2 ...
##  $ area_worst             : num  2019 1956 1709 568 1575 ...
##  $ smoothness_worst       : num  0.162 0.124 0.144 0.21 0.137 ...
##  $ compactness_worst      : num  0.666 0.187 0.424 0.866 0.205 ...
##  $ concavity_worst        : num  0.712 0.242 0.45 0.687 0.4 ...
##  $ concave.points_worst   : num  0.265 0.186 0.243 0.258 0.163 ...
##  $ symmetry_worst         : num  0.46 0.275 0.361 0.664 0.236 ...
##  $ fractal_dimension_worst: num  0.1189 0.089 0.0876 0.173 0.0768 ...
##  $ X                      : logi  NA NA NA NA NA NA ...
```

```
head(df)
```

id	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]

```
summary(df)
```

```
##        id             diagnosis  radius_mean      texture_mean
##  Min.   :     8670   B:357     Min.   : 6.981   Min.   : 9.71
##  1st Qu.:   869218   M:212     1st Qu.:11.700   1st Qu.:16.17
##  Median :   906024             Median :13.370   Median :18.84
##  Mean   : 30371831             Mean   :14.127   Mean   :19.29
##  3rd Qu.:  8813129             3rd Qu.:15.780   3rd Qu.:21.80
##  Max.   :911320502             Max.   :28.110   Max.   :39.28
##  perimeter_mean     area_mean      smoothness_mean   compactness_mean
##  Min.   : 43.79   Min.   : 143.5   Min.   :0.05263   Min.   :0.01938
##  1st Qu.: 75.17   1st Qu.: 420.3   1st Qu.:0.08637   1st Qu.:0.06492
##  Median : 86.24   Median : 551.1   Median :0.09587   Median :0.09263
##  Mean   : 91.97   Mean   : 654.9   Mean   :0.09636   Mean   :0.10434
##  3rd Qu.:104.10   3rd Qu.: 782.7   3rd Qu.:0.10530   3rd Qu.:0.13040
##  Max.   :188.50   Max.   :2501.0   Max.   :0.16340   Max.   :0.34540
##  concavity_mean    concave.points_mean symmetry_mean    fractal_dimension_m
##  Min.   :0.00000   Min.   :0.00000     Min.   :0.1060   Min.   :0.04996
##  1st Qu.:0.02956   1st Qu.:0.02031     1st Qu.:0.1619   1st Qu.:0.05770
##  Median :0.06154   Median :0.03350     Median :0.1792   Median :0.06154
##  Mean   :0.08880   Mean   :0.04892     Mean   :0.1812   Mean   :0.06280
##  3rd Qu.:0.13070   3rd Qu.:0.07400     3rd Qu.:0.1957   3rd Qu.:0.06612
##  Max.   :0.42680   Max.   :0.20120     Max.   :0.3040   Max.   :0.09744
##    radius_se        texture_se      perimeter_se       area_se
##  Min.   :0.1115   Min.   :0.3602   Min.   : 0.757   Min.   :  6.802
##  1st Qu.:0.2324   1st Qu.:0.8339   1st Qu.: 1.606   1st Qu.: 17.850
##  Median :0.3242   Median :1.1080   Median : 2.287   Median : 24.530
##  Mean   :0.4052   Mean   :1.2169   Mean   : 2.866   Mean   : 40.337
##  3rd Qu.:0.4789   3rd Qu.:1.4740   3rd Qu.: 3.357   3rd Qu.: 45.190
##  Max.   :2.8730   Max.   :4.8850   Max.   :21.980   Max.   :542.200
##  smoothness_se      compactness_se      concavity_se     concave.points_se
##  Min.   :0.001713   Min.   :0.002252   Min.   :0.00000   Min.   :0.000000
##  1st Qu.:0.005169   1st Qu.:0.013080   1st Qu.:0.01509   1st Qu.:0.007638
##  Median :0.006380   Median :0.020450   Median :0.02589   Median :0.010930
##  Mean   :0.007041   Mean   :0.025478   Mean   :0.03189   Mean   :0.011796
##  3rd Qu.:0.008146   3rd Qu.:0.032450   3rd Qu.:0.04205   3rd Qu.:0.014710
##  Max.   :0.031130   Max.   :0.135400   Max.   :0.39600   Max.   :0.052790
##   symmetry_se       fractal_dimension_se  radius_worst   texture_worst
##  Min.   :0.007882   Min.   :0.0008948    Min.   : 7.93   Min.   :12.02
##  1st Qu.:0.015160   1st Qu.:0.0022480    1st Qu.:13.01   1st Qu.:21.08
##  Median :0.018730   Median :0.0031870    Median :14.97   Median :25.41
##  Mean   :0.020542   Mean   :0.0037949    Mean   :16.27   Mean   :25.68
```

```
##  3rd Qu.:0.023480   3rd Qu.:0.0045580    3rd Qu.:18.79   3rd Qu.:29.72
##  Max.   :0.078950   Max.   :0.0298400    Max.   :36.04   Max.   :49.54
##  perimeter_worst    area_worst     smoothness_worst  compactness_worst
##  Min.   : 50.41   Min.   : 185.2   Min.   :0.07117   Min.   :0.02729
##  1st Qu.: 84.11   1st Qu.: 515.3   1st Qu.:0.11660   1st Qu.:0.14720
##  Median : 97.66   Median : 686.5   Median :0.13130   Median :0.21190
##  Mean   :107.26   Mean   : 880.6   Mean   :0.13237   Mean   :0.25427
##  3rd Qu.:125.40   3rd Qu.:1084.0   3rd Qu.:0.14600   3rd Qu.:0.33910
##  Max.   :251.20   Max.   :4254.0   Max.   :0.22260   Max.   :1.05800
##  concavity_worst  concave.points_worst symmetry_worst   fractal_dimension_w
##  Min.   :0.0000   Min.   :0.00000      Min.   :0.1565   Min.   :0.05504
##  1st Qu.:0.1145   1st Qu.:0.06493      1st Qu.:0.2504   1st Qu.:0.07146
##  Median :0.2267   Median :0.09993      Median :0.2822   Median :0.08004
##  Mean   :0.2722   Mean   :0.11461      Mean   :0.2901   Mean   :0.08395
##  3rd Qu.:0.3829   3rd Qu.:0.16140      3rd Qu.:0.3179   3rd Qu.:0.09208
##  Max.   :1.2520   Max.   :0.29100      Max.   :0.6638   Max.   :0.20750
##     X
##  Mode:logical
##  NA's:569
##
##
##
##
```

There are a lot of variables, we should pick the most relevant ones.

Let's delete the last variable and ID number because there are not relevant.

```
df<-df[,-33]
df<-df[,-1]
```

Proportion of benign vs malignant cancer

```
prop.table(table(df$diagnosis))    #  B : 0.6274165  M: 0.3725835
```

```
##
##         B         M
## 0.6274165 0.3725835
```

The two types of cancer are not represented in the same proportion, this can lead to a bias. Is this proportion representative of the reality ?

> "The benign to malignant ratio (B:M ratio) among breast biopsies (number of benign breast lesions divided by number of breast cancers) is widely believed to be around 4:1 or 5:1" [2]

2.3 2. Selection of variables of interest

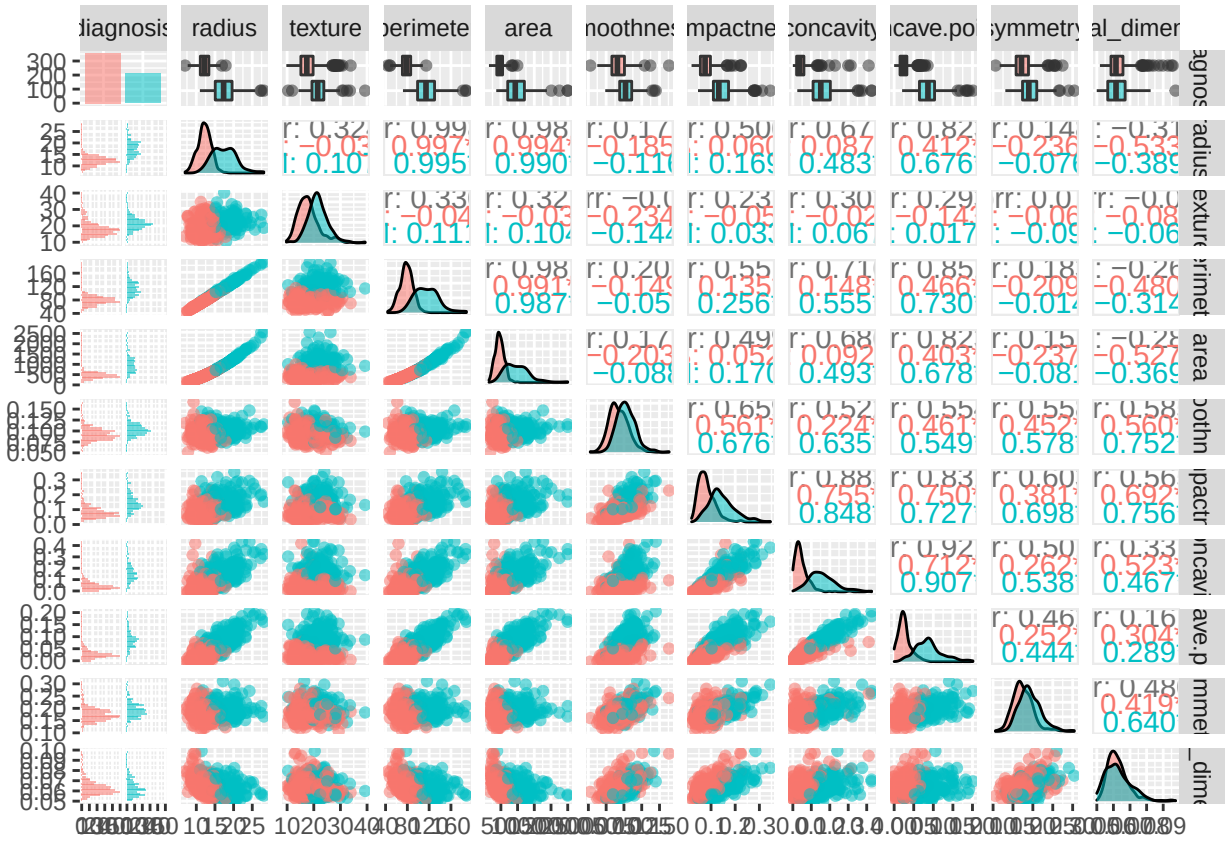
According to the description of the data, some variables are likely to be correlated. We will address this correlation with the mean group of variable.

Hypothesis about the correlation between variables : *radius and smoothness should be correlated* radius, perimeter, area and compactness should be perfectly correlated since it exists a formula between these variables * concavity and symmetry should be correlated * texture and fractal_dimension should not have any correlation

Let's create a new frame for a the variables of type mean.

```
## 'data.frame':    569 obs. of  11 variables:
##  $ diagnosis        : Factor w/ 2 levels "B","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ radius           : num  18 20.6 19.7 11.4 20.3 ...
##  $ texture          : num  10.4 17.8 21.2 20.4 14.3 ...
##  $ perimeter        : num  122.8 132.9 130 77.6 135.1 ...
##  $ area             : num  1001 1326 1203 386 1297 ...
##  $ smoothness       : num  0.1184 0.0847 0.1096 0.1425 0.1003 ...
##  $ compactness      : num  0.2776 0.0786 0.1599 0.2839 0.1328 ...
##  $ concavity        : num  0.3001 0.0869 0.1974 0.2414 0.198 ...
##  $ concave.points   : num  0.1471 0.0702 0.1279 0.1052 0.1043 ...
##  $ symmetry         : num  0.242 0.181 0.207 0.26 0.181 ...
##  $ fractal_dimension: num  0.0787 0.0567 0.06 0.0974 0.0588 ...
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## [1] "B" "M"
```

By eye, the variables seem to be different according to the type of 'diagnosis' (first row of the plot), the variables coming from malignant cancer seem to be in general bigger than the data coming from bening cancer.

As expected, radius, perimeter and area are highly correlated;
and texture and fractal_dimension don't have strong correlation.

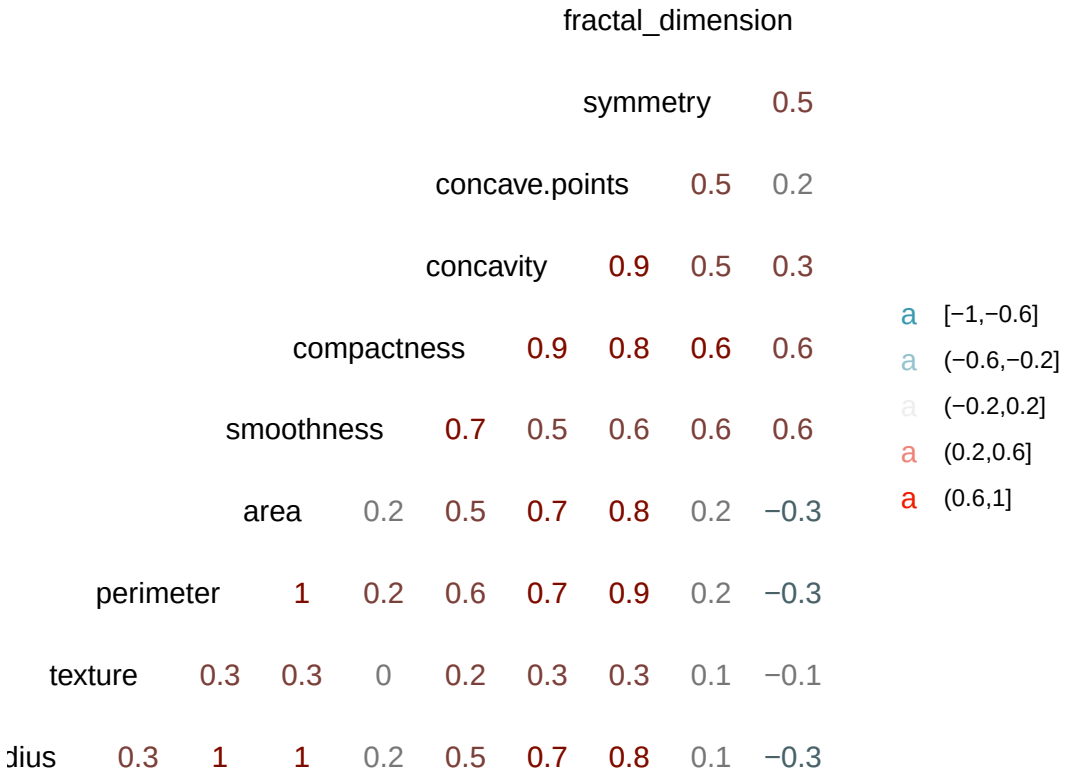
Surprisingly, radius and smoothness are not very correlated, and the compactness doesn't show any strong correlation.

Concavity and compactness have a strong correlation. In annexe we show that we have the same correlations with the standard deviation group and extreme group of variable.

The following function permit to see better the correlation :

```
ggcorr(df_mean, geom = "text", nbreaks = 5, hjust = 1, label = TRUE, label_al
```

```
## Warning in ggcorr(df_mean, geom = "text", nbreaks = 5, hjust = 1, label =
## TRUE, : data in column(s) 'diagnosis' are not numeric and were ignored
```



Even if compactness is define as $perimeter^2/area - 1.0$ the correlation between this variable and area or perimeter is not 1 because the correlation show only the linear dependency. The correlation between radius perimeter and area is 1. We will only keep radius.

From here let's remove area, perimeter and compactness

```
m_perimeter <-lm(data = df_mean,perimeter~ radius+texture+area+smoothness+comp
summary(m_perimeter)
```

```
##
## Call:
## lm(formula = perimeter ~ radius + texture + area + smoothness +
##     compactness + concavity + concave.points + symmetry + fractal_dimensio
##     data = df_mean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7781 -0.1859  0.0416  0.2280  3.8482
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.627e+00  8.317e-01   3.159  0.00167 **
## radius            6.126e+00  5.223e-02 117.301  < 2e-16 ***
## texture          -2.105e-03  5.886e-03  -0.358  0.72073
```

```
## area                3.868e-03  4.676e-04   8.272 9.65e-16 ***
## smoothness         -8.998e+00  2.816e+00  -3.196  0.00147 **
## compactness         3.407e+01  1.517e+00  22.459  < 2e-16 ***
## concavity           3.865e+00  9.842e-01   3.927 9.67e-05 ***
## concave.points      4.283e+00  2.785e+00   1.538  0.12456
## symmetry           -1.930e+00  1.127e+00  -1.712  0.08751 .
## fractal_dimension -4.119e+01  8.190e+00  -5.029 6.64e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5538 on 559 degrees of freedom
## Multiple R-squared:  0.9995, Adjusted R-squared:  0.9995
## F-statistic: 1.214e+05 on 9 and 559 DF,  p-value: < 2.2e-16
```

```
m_area <-lm(data = df_mean,area~ radius+texture+ perimeter+smoothness+compactn
summary(m_area)
```

```
##
## Call:
## lm(formula = area ~ radius + texture + perimeter + smoothness +
##     compactness + concavity + concave.points + symmetry + fractal_dimensio
##     data = df_mean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -119.45  -25.53   -8.03   18.13  395.17
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1033.1844    56.7713 -18.199  < 2e-16 ***
## radius              -81.3418    22.3034  -3.647  0.00029 ***
## texture               0.4089     0.5023   0.814  0.41588
## perimeter            28.1972     3.4086   8.272 9.65e-16 ***
## smoothness           92.0982   242.5499   0.380  0.70431
## compactness       -2169.2200   153.3091 -14.149  < 2e-16 ***
## concavity           221.0651    84.6672   2.611  0.00927 **
## concave.points      295.3508   237.9177   1.241  0.21498
## symmetry             92.1348    96.4335   0.955  0.33978
## fractal_dimension  6413.4180   661.4625   9.696  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47.28 on 559 degrees of freedom
## Multiple R-squared:  0.9822, Adjusted R-squared:  0.9819
## F-statistic:  3434 on 9 and 559 DF,  p-value: < 2.2e-16
```

```
m_compactness <-lm(data = df_mean, compactness~ radius+texture+ perimeter+smoo
summary(m_compactness)
```

```
##
## Call:
## lm(formula = compactness ~ radius + texture + perimeter + smoothness +
##     area + concavity + concave.points + symmetry + fractal_dimension,
##     data = df_mean)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.044976 -0.006177 -0.000675  0.005421  0.058426
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -2.053e-01  1.457e-02 -14.091  < 2e-16 ***
## radius            -7.702e-02  4.234e-03 -18.189  < 2e-16 ***
## texture            2.376e-04  1.185e-04   2.004  0.04555 *
## perimeter          1.392e-02  6.198e-04  22.459  < 2e-16 ***
## smoothness         1.759e-01  5.694e-02   3.089  0.00211 **
## area              -1.216e-04  8.592e-06 -14.149  < 2e-16 ***
## concavity          6.410e-02  1.998e-02   3.208  0.00141 **
## concave.points     1.077e-01  5.622e-02   1.915  0.05598 .
## symmetry           9.448e-02  2.250e-02   4.200  3.1e-05 ***
## fractal_dimension  2.348e+00  1.370e-01  17.136  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01119 on 559 degrees of freedom
## Multiple R-squared:  0.9558, Adjusted R-squared:  0.9551
## F-statistic:  1343 on 9 and 559 DF,  p-value: < 2.2e-16
```

The variables area, perimeter and compactness are well explained by the other variables (Adjusted R-squared very close to 1). So we discard them.

3 GLM

```
m <-glm(data = df_mean, diagnosis~ radius+texture+smoothness+concavity+concave
summary(m)
```

```
##
## Call:
## glm(formula = diagnosis ~ radius + texture + smoothness + concavity +
##     concave.points + symmetry + fractal_dimension, family = binomial,
```

```
##     data = df_mean)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.35180  -0.13938  -0.03229   0.02046   3.15368
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -28.38387    6.66946  -4.256 2.08e-05 ***
## radius               0.88701    0.21852   4.059 4.93e-05 ***
## texture              0.37262    0.06212   5.998 2.00e-09 ***
## smoothness          78.50170   32.64920   2.404   0.0162 *
## concavity           15.52082    8.35462   1.858   0.0632 .
## concave.points      46.67203   26.16265   1.784   0.0744 .
## symmetry            16.85783   10.75613   1.567   0.1170
## fractal_dimension -101.54448   61.26233  -1.658   0.0974 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 751.44  on 568  degrees of freedom
## Residual deviance: 153.35  on 561  degrees of freedom
## AIC: 169.35
##
## Number of Fisher Scoring iterations: 8
```

3.1 Validation

```
751.44/153.35
```

```
## [1] 4.900163
```

Overdistribution ?

Let's see with an anova test if we can remove symmetry,concavity,concave.points and fractal_dimension

```
m1 <-glm(data = df_mean, diagnosis~ radius+texture+smoothness+concavity+concav
summary(m1)
```

```
##
## Call:
## glm(formula = diagnosis ~ radius + texture + smoothness + concavity +
##     concave.points + fractal_dimension, family = binomial, data = df_mean)
##
```

```
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.33122  -0.15084  -0.03480   0.02274   3.04740
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -26.65706    6.59839  -4.040 5.35e-05 ***
## radius              0.86784    0.22209   3.908 9.32e-05 ***
## texture             0.36277    0.06098   5.949 2.70e-09 ***
## smoothness         90.53604   33.17393   2.729  0.00635 **
## concavity          17.34487    8.23293   2.107  0.03514 *
## concave.points     45.65526   26.56395   1.719  0.08567 .
## fractal_dimension -93.18039   59.52661  -1.565  0.11750
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 751.44  on 568  degrees of freedom
## Residual deviance: 155.78  on 562  degrees of freedom
## AIC: 169.78
##
## Number of Fisher Scoring iterations: 8
```

```
anova(m,m1,test="Chisq")
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
561	153.3485	NA	NA	NA
562	155.7755	-1	-2.426971	0.1192631

```
m2 <-glm(data = df_mean, diagnosis~ radius+texture+smoothness+concavity+concav
summary(m2)
```

```
##
## Call:
## glm(formula = diagnosis ~ radius + texture + smoothness + concavity +
##     concave.points, family = binomial, data = df_mean)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.28927  -0.15267  -0.03761   0.02390   3.03440
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -32.18048    5.69324  -5.652 1.58e-08 ***
```

```
## radius           0.99766    0.20762   4.805 1.55e-06 ***
## texture          0.36496    0.06131   5.953 2.64e-09 ***
## smoothness      72.72278   30.46830   2.387   0.0170 *
## concavity       10.21913    6.99120   1.462   0.1438
## concave.points  48.66262   26.54605   1.833   0.0668 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 751.44  on 568  degrees of freedom
## Residual deviance: 158.34  on 563  degrees of freedom
## AIC: 170.34
##
## Number of Fisher Scoring iterations: 8
```

```
anova(m1,m2,test="Chisq")
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
562	155.7755	NA	NA	NA
563	158.3372	-1	-2.56172	0.1094794

```
m3 <-glm(data = df_mean, diagnosis~ radius+texture+smoothness+concave.points,f
summary(m3)
```

```
##
## Call:
## glm(formula = diagnosis ~ radius + texture + smoothness + concave.points,
##     family = binomial, data = df_mean)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.42132  -0.15010  -0.04247   0.02603   2.86598
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -28.57552    4.81406  -5.936 2.92e-09 ***
## radius           0.85081    0.17112   4.972 6.63e-07 ***
## texture          0.35845    0.05985   5.990 2.10e-09 ***
## smoothness      52.26403   26.08496   2.004   0.0451 *
## concave.points  78.73692   16.59332   4.745 2.08e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 751.44  on 568  degrees of freedom
## Residual deviance: 160.32  on 564  degrees of freedom
## AIC: 170.32
##
## Number of Fisher Scoring iterations: 8
```

```
anova(m2,m3,test="Chisq")
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
563	158.3372	NA	NA	NA
564	160.3203	-1	-1.983059	0.1590686

3.2 Visualization

4 PCA

---

5 Annexe

5.1 More informations about the dataset

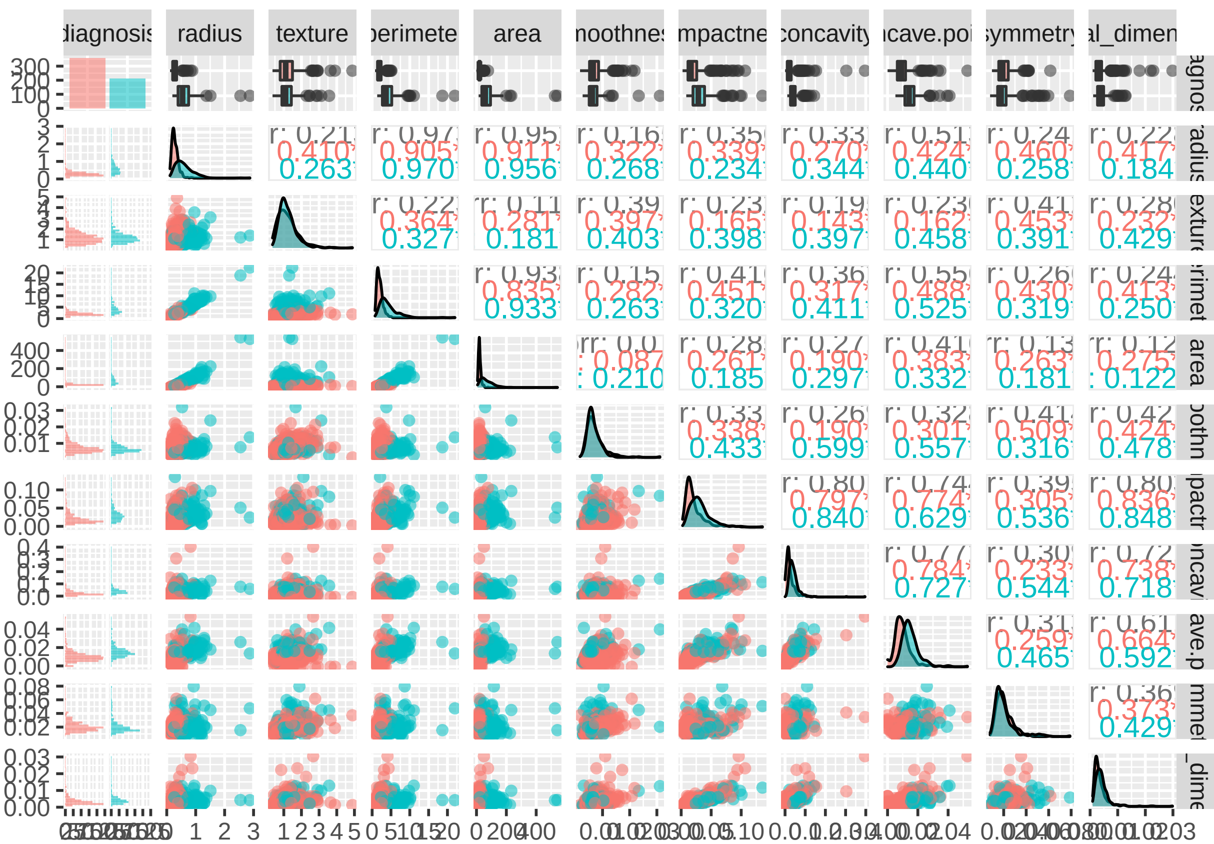
The 3-dimensional space is that described in: [3].

This database is also available through the UW CS ftp server: ftp ftp.cs.wisc.edu cd math-prog/cpo-dataset/machine-learn/WDBC/

5.2 Correlation in the 'standard deviation' and 'worst' group

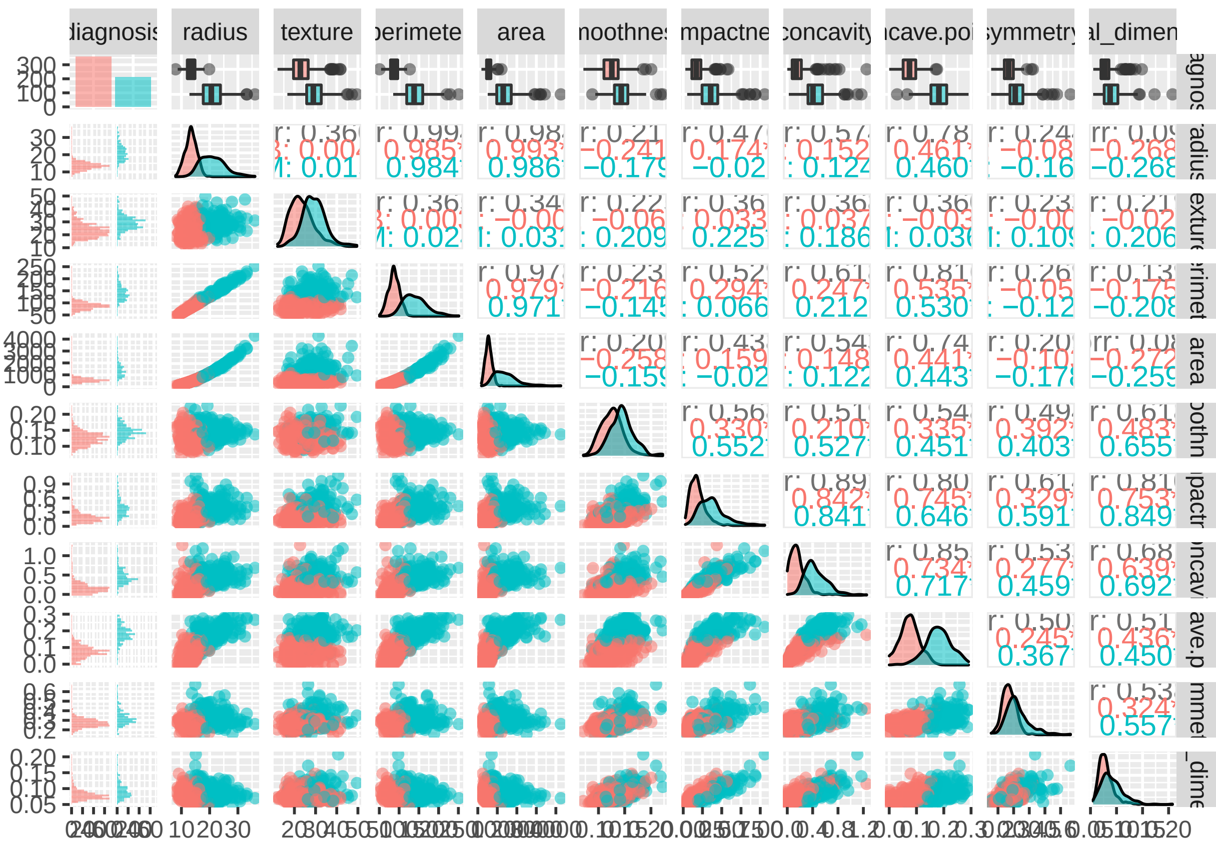
```
## 'data.frame':    569 obs. of  11 variables:
##  $ diagnosis         : Factor w/ 2 levels "B","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ radius            : num  1.095 0.543 0.746 0.496 0.757 ...
##  $ texture           : num  0.905 0.734 0.787 1.156 0.781 ...
##  $ perimeter         : num  8.59 3.4 4.58 3.44 5.44 ...
##  $ area              : num  153.4 74.1 94 27.2 94.4 ...
##  $ smoothness        : num  0.0064 0.00522 0.00615 0.00911 0.01149 ...
##  $ compactness       : num  0.049 0.0131 0.0401 0.0746 0.0246 ...
##  $ concavity         : num  0.0537 0.0186 0.0383 0.0566 0.0569 ...
##  $ concave.points    : num  0.0159 0.0134 0.0206 0.0187 0.0188 ...
##  $ symmetry          : num  0.03 0.0139 0.0225 0.0596 0.0176 ...
##  $ fractal_dimension.: num  0.00619 0.00353 0.00457 0.00921 0.00511 ...
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## 'data.frame':    569 obs. of  11 variables:
##  $ diagnosis          : Factor w/ 2 levels "B","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ radius             : num  25.4 25 23.6 14.9 22.5 ...
##  $ texture            : num  17.3 23.4 25.5 26.5 16.7 ...
##  $ perimeter          : num  184.6 158.8 152.5 98.9 152.2 ...
##  $ area               : num  2019 1956 1709 568 1575 ...
##  $ smoothness         : num  0.162 0.124 0.144 0.21 0.137 ...
##  $ compactness        : num  0.666 0.187 0.424 0.866 0.205 ...
##  $ concavity          : num  0.712 0.242 0.45 0.687 0.4 ...
##  $ concave.points     : num  0.265 0.186 0.243 0.258 0.163 ...
##  $ symmetry           : num  0.46 0.275 0.361 0.664 0.236 ...
##  $ fractal_dimension.: num  0.1189 0.089 0.0876 0.173 0.0768 ...
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We get the same results as for the mean group.

---

6 References

- [1] https://www.researchgate.net/figure/a-b-Fine-needle-aspiration-cytology-of-the-breast-lesion-showed-singly-lying_fig1_41548857
- [2] https://pubmed.ncbi.nlm.nih.gov/7091922/
- [3] K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34

7 Version of R used

```
## R version 4.1.1 (2021-08-10)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRblas
## LAPACK: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRlapa
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] MASS_7.3-54      ggfortify_0.4.12 GGally_2.1.2     ggplot2_3.3.5
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.7         highr_0.9          pillar_1.6.3       compiler_4.1
##  [5] RColorBrewer_1.1-2 plyr_1.8.6         tools_4.1.1        digest_0.6.28
##  [9] evaluate_0.14      lifecycle_1.0.1    tibble_3.1.5       gtable_0.3.0
## [13] pkgconfig_2.0.3    rlang_0.4.11       yaml_2.2.1         xfun_0.26
## [17] fastmap_1.1.0      gridExtra_2.3      withr_2.4.2        stringr_1.4.0
## [21] dplyr_1.0.7        knitr_1.36         generics_0.1.0     vctrs_0.3.8
## [25] grid_4.1.1         tidyselect_1.1.1   reshape_0.8.8      glue_1.4.2
## [29] R6_2.5.1           fansi_0.5.0        rmarkdown_2.11     farver_2.1.0
## [33] tidyr_1.1.4        purrr_0.3.4        magrittr_2.0.1     scales_1.1.1
## [37] ellipsis_0.3.2     htmltools_0.5.2    colorspace_2.0-2   labeling_0.4
## [41] utf8_1.2.2         stringi_1.7.5      munsell_0.5.0      crayon_1.4.1
```