Mini Project

DSC4033 - Multivariate Methods II

S.Thineskumar
(S/19/582)

Department of Statistics and Computer Science
Faculty of Science
University of Peradeniya

2025

Contents

1 Introduction

Breast cancer is a disease in which abnormal breast cells grow out of control and form tumours. If left unchecked, the tumours can spread throughout the body and become fatal. Breast cancer cells begin inside the milk ducts and/or the milk-producing lobules of the breast. Cancer cells can spread into nearby breast tissue (invasion). This creates tumours that cause lumps or thickening. The earliest form (in situ) is not life-threatening and can be detected in early stages. Invasive cancers can spread to nearby lymph nodes or other organs (metastasize). Metastasis can be life-threatening and fatal. In 2022, there were 2.3 million women diagnosed with breast cancer and 670 000 deaths globally. Breast cancer occurs in every country of the world in women at any age after puberty but with increasing rates in later life (WHO). Global estimates reveal striking inequities in the breast cancer burden according to human development. For instance, in countries with a very high Human Development Index (HDI), 1 in 12 women will be diagnosed with breast cancer in their lifetime and 1 in 71 women die of it. So, we are in a situation that figure out the stage of the breast cancer. In this project we focus on predict the stage of the breast cancer and analysing the causes of breast cancer.

2 Methodology

2.1 Dataset Description

The Breast Cancer dataset downloaded from the UC Irvine Machine Learning Repository. The data set contains 30 features and 569 instances (rows). 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. The features of dataset presented in Table 1.

Table 1: Feature descriptions for breast cancer dataset

Features	**Role**	**Type**	**Description**
ID	ID	Categorical	ID number
Diagnosis	Target	Categorical	M = malignant, B = benign
radius	Feature	Continuous	Mean of distances from center to points on the perimeter
texture	Feature	Continuous	Standard deviation of gray-scale values
perimeter	Feature	Continuous	Length of the contour
area	Feature	Continuous	Area inside the contour
smoothness	Feature	Continuous	Local variation in radius lengths
compactness	Feature	Continuous	$\frac{\text{perimeter}^2}{\text{area}} - 1.0$
concavity	Feature	Continuous	Severity of concave portions of the contour
concave_points	Feature	Continuous	Number of concave portions of the contour
symmetry	Feature	Continuous	Symmetry of the shape
fractal_dimension	Feature	Continuous	Coastline approximation - 1

The actual linear program used to obtain the separating plane in the 3-dimensional space. So, we got 3 sets of radius, texture, perimeter, area, smoothness, compactness, concavity, concave_points, symmetry, fractal_dimension variables.

2.2 Data Preprocessing

- **Missing Value**
  Check the columns contain any missing value. If So, fill them with a appropriate vales (mean, median, nearest neighbours).

- **Standardized the numerical values**
  Variables used in multivariate techniques (CCA , PCA, SEM) such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave_points, symmetry, fractal_dimension were standardized using the scale() function.

2.3 Statistical Techniques

- **Principal Component Analysis**
  Principal Component Analysis (PCA) is a statistical technique used for dimensionality reduction and data visualization. PCA aims to transform a high-dimensional dataset into a lower-dimensional space while retaining the most important information.The main objective of PCA is to find a set of new variables, called principal components, that capture the maximum variance in the original dataset. These components are obtained as linear combinations of the original variables.The first principal component explains the largest amount of variance in the data, followed by the second component, and so on. Each subsequent component is uncorrelated to the previous components.

- **Factor Analysis**
  Factor Analysis is a statistical technique used to identify underlying factors or latent variables that explain the correlations among a set of observed variables. The main goal of factor analysis is to determine the number of factors and understand how each variable relates to these factors. It assumes that each observed variable is influenced by one or more latent factors

- **Discriminant Analysis**
  Discriminant analysis is a power full statistical technique used to classify observations into predefined groups based on predictor variables. It is particularly useful when dealing with classification problem where the response variable is categorical. Mainly we have two types of discriminant analysis. They are Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). Based on the Bartlett's test we decide the Discriminant analysis type.

- **Canonical Correlation Analysis**
  Canonical correlation analysis is a method for exploring the relationship between two multivariate sets of variables, all measured on the same individual. Canonical correlation analysis allows us to summarize the relationship into lesser number of statistics while preserving the main facets of the relationships. This is another dimension reduction technique. These are the main objectives of Canonical Correlation Analysis. Test the hypothesis that canonical variate pairs are correlated or not (canonical correlations are equal or not equal to zero) at 1% significance level.

- **Structural Equation Modelling**
  Structural Equation Modelling (SEM) is a comprehensive statistical framework that incorporates both measurement models and structural models. SEM extend the concept of covariance structure models by including latent variables. which are unobserved or underlying constructs that cannot be measured directly.

3 Results and Discussion

3.1 Principal Component Analysis

- **Correlation matrix**

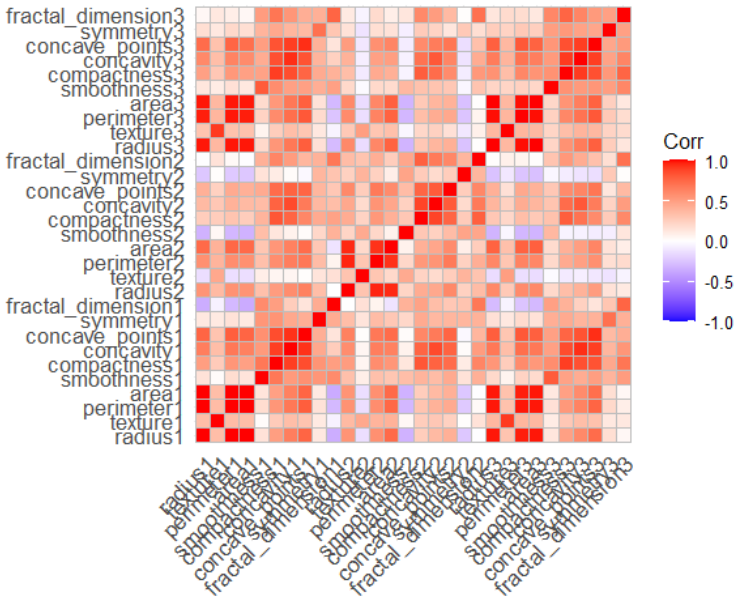


Figure 1: Correlation Matrix

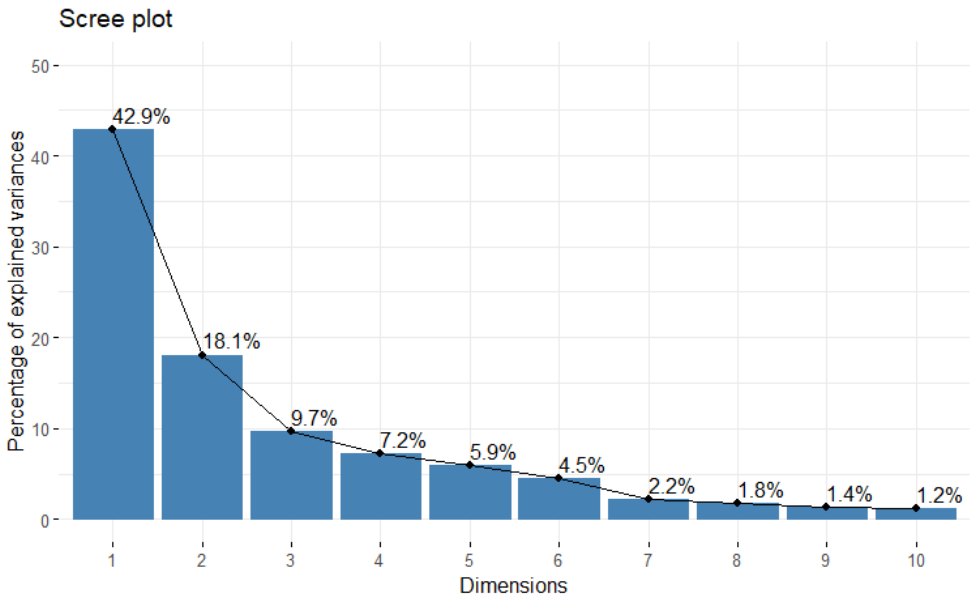
All three radius variable has highest correlation with perimeter, and area variables. So, we need to remove them.

- **Eigen values**

Table 2: Eigen values for PCA

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
Values	11.5808	4.8810	2.6217	1.9500	1.6005	1.2062	0.6048	0.4753	0.3659	0.3288
	λ_{11}	λ_{12}	λ_{13}	λ_{14}	λ_{15}	λ_{16}	λ_{17}	λ_{18}	λ_{19}	λ_{20}
Values	0.2865	0.2584	0.2344	0.1561	0.0901	0.0759	0.0564	0.0453	0.0419	0.0309
	λ_{21}	λ_{22}	λ_{23}	λ_{24}	λ_{25}	λ_{26}	λ_{27}			
Values	0.0295	0.0259	0.0231	0.0154	0.0079	0.0061	0.0012			

- **Scree plot**



According to the Jollieffe's Method and scree plot. We select first 6 principal components to represent the dataset adequetly.

3.2 Factor Analysis

- **Kaiser-Meyer-Olkin Test (KMO Test)**

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = scaled_breast_cancer)
Overall MSA =  0.79
MSA for each item =
          texture1         perimeter1              area1         smoothness1
              0.61               0.73               0.77               0.81
     compactness1        concavity1    concave_points1        symmetry1
              0.89               0.87               0.88               0.82
fractal_dimension1           texture2         perimeter2              area2
              0.80               0.46               0.70               0.73
      smoothness2       compactness2        concavity2    concave_points2
              0.62               0.88               0.83               0.81
        symmetry2 fractal_dimension2           texture3         perimeter3
              0.56               0.83               0.58               0.72
           area3        smoothness3       compactness3        concavity3
              0.71               0.75               0.85               0.89
  concave_points3          symmetry3 fractal_dimension3
              0.87               0.68               0.81
```

Remove all low contribute (MSAi <0.5) variable to increase the MSA value. Here we remove the texture2 variable.

- **Bartlett's Test**

H_0 : Correlation matrix is an identity matrix Vs. H_1 : Correlation matrix is different from an identity matrix

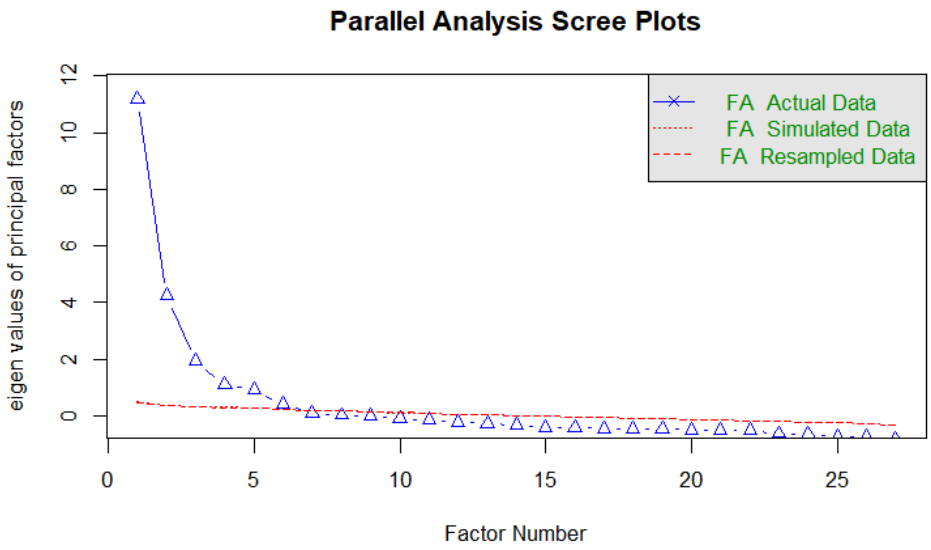
```
R was not square, finding R from data
$chisq
[1] 28432.21

$p.value
[1] 0

$df
[1] 325
```

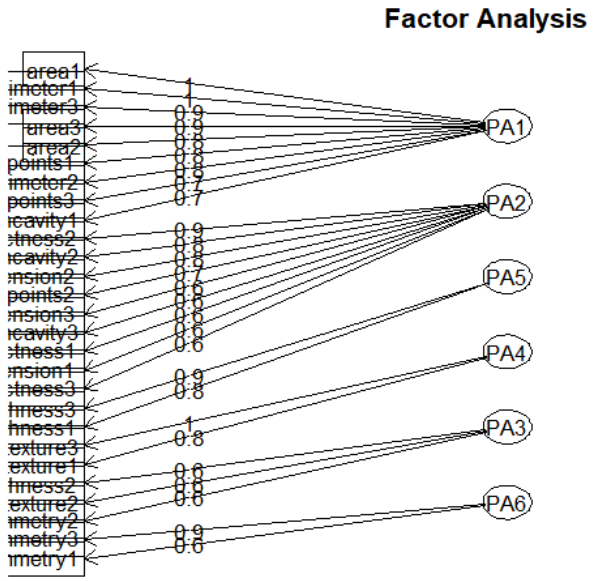
p-value < 0.05. So, we reject H0. We can conclude that correlation matrix is significantly different from an identity matrix at 5% significant level. Suggesting that factor Analysis is appropriate.

- **Scree plot and Parallel analysis scree plot**



Parallel analysis suggests that the number of factors is equal to 6.

- **Graph factor loadings**



3.3 Discriminant Analysis

- **Bartlett's Test**

  Use Bartlett's Test to determine if variance-covariance matrices are homogeneous for the two or more populations involved.

```
        Bartlett test of homogeneity of variances

data:  radius1 by Diagnosis
Bartlett's K-squared = 95.429, df = 1, p-value < 2.2e-16


        Bartlett test of homogeneity of variances

data:  texture1 by Diagnosis
Bartlett's K-squared = 0.80628, df = 1, p-value = 0.3692


        Bartlett test of homogeneity of variances

data:  perimeter1 by Diagnosis
Bartlett's K-squared = 104.74, df = 1, p-value < 2.2e-16


        Bartlett test of homogeneity of variances

data:  area1 by Diagnosis
Bartlett's K-squared = 271.59, df = 1, p-value < 2.2e-16


        Bartlett test of homogeneity of variances

data:  smoothness1 by Diagnosis
Bartlett's K-squared = 1.082, df = 1, p-value = 0.2982
```

Bartlett's test p-value for texture1, smoothness1, symmetry1, fractal_dimension1, smoothness2, compactnesss2, concave_points2, texture3, smoothness3 are greater than 0.05. p-value for other variables are <0.05. So, we reject the H_0. Suggest that the variance-covariance matrix are different across the groups.So, we will use Quadratic Discriminant Analysis.

- **Confusion Matrix for test data**

	Benign	Malignant
Benign	52	1
Malignant	2	42

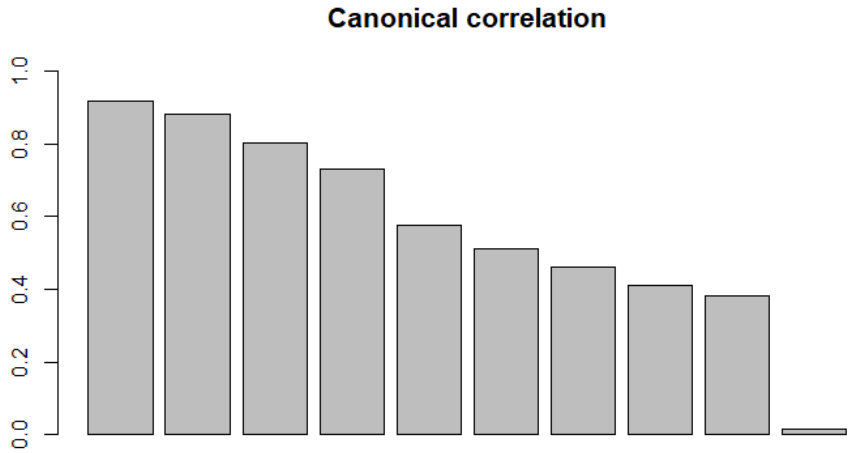
Test accuracy is 96.90%. The misclassified percentage is very small. Our model predict the variable very well.

3.4 Canonical Correlation Analysis

In canonical correlation analysis we compare the values between two sets of plane.
Set1(X) - radius1, texture1, perimeter1, area1, smoothness1, compactness1, concavity1, concave_points1, symmetry1, fractal_dimension1
Set2(Y) - radius2, texture2, perimeter2, area2, smoothness2, compactness2, concavity2, concave_points2, symmetry2, fractal_dimension2.

- **Canonical Correlation**

0.9175	0.8810	0.8019	0.7301	0.5770	0.5132	0.4630	0.4106	0.3822	0.0173



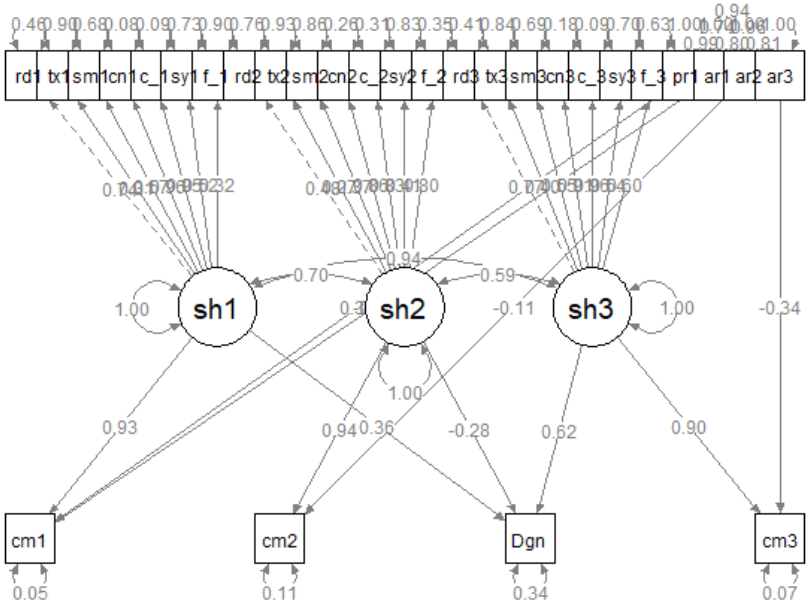
- **Significant test for canonical correlation**

```
Wilks' Lambda, using F-approximation (Rao's F):
                stat     approx df1      df2   p.value
1 to 10:   0.001617355 57.3184397 100 3942.542 0.0000000
2 to 10:   0.010227837 45.4078082  81 3563.362 0.0000000
3 to 10:   0.045678776 35.2066312  64 3184.586 0.0000000
4 to 10:   0.127953170 28.6007705  49 2806.832 0.0000000
5 to 10:   0.273994938 23.1557100  36 2431.154 0.0000000
6 to 10:   0.410763013 22.2943877  25 2059.521 0.0000000
7 to 10:   0.557647970 22.3327489  16 1696.191 0.0000000
8 to 10:   0.709778095 22.7441773   9 1353.309 0.0000000
9 to 10:   0.853690489 22.9220214   4 1114.000 0.0000000
10 to 10:  0.999701835  0.1664258   1  558.000 0.6834645
```

The first test of the canonical dimensions tests whether all 10 dimensions are significant, the next test tests whether dimension 2 to 10 are significant so on,the last test tests whether dimension 10 by itself, is significant. All p-value are less than 0.05 other than last one. So, we can conclude that the all other dimensions are significant except dimension 10.
We have 9 canonical covariate pairs.

3.5 Structural Equation Modelling

- **Structural Paths**



we create 3 latent variables (Shape) using 3 sets of radius, texture, smoothness, compactness, concavity, concave_points, symmetry, fractal_dimension. Using these laten variables we will predict the diagnosis variable.

- **SEM model**

```
lavaan 0.6.17 ended normally after 69 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        59

  Number of observations                           569

Model Test User Model:

  Test statistic                             21309.389
  Degrees of freedom                               366
  P-value (Chi-square)                           0.000

Model Test Baseline Model:

  Test statistic                             31258.120
  Degrees of freedom                               400
  P-value                                        0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.321
  Tucker-Lewis Index (TLI)                       0.258

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)             -14784.521
  Loglikelihood unrestricted model (H1)      -4129.827

  Akaike (AIC)                               29687.043
  Bayesian (BIC)                             29943.332
  Sample-size adjusted Bayesian (SABIC)      29756.033

Root Mean Square Error of Approximation:

  RMSEA                                          0.317
  90 Percent confidence interval - lower         0.314
  90 Percent confidence interval - upper         0.321
  P-value H_0: RMSEA <= 0.050                    0.000
  P-value H_0: RMSEA >= 0.080                    1.000

Standardized Root Mean Square Residual:

  SRMR                                           0.304

Parameter Estimates:

  Standard errors                             Standard
  Information                                 Expected
  Information saturated (h1) model          Structured
```

The model demonstrated poor fit based on RMSEA = 0.317, CFI = 0.321, TLI = 0.258, and SRMR = 0.304, indicating that improvements in model specification may be necessary.

4 Conclusion and Recommendation

This study utilised a well-established dataset and a variety of multivariate techniques to analyse and predict the stages of breast cancer. While Discriminant Analysis created a high classification accuracy of 96.9%, indicating its efficacy in distinct between benign and malignant cases, Principal Component Analysis and Factor Analysis helped in reducing dimensionality and identifying underlying variable structures. Strong correlations between various sets of variables were found using canonical correlation analysis. The Structural Equation Model showed poor overall model fit (RMSEA = 0.317), indicating the need for further model refinement, despite successfully validating important theoretical relationships, such as the derived compactness formula. We recommend making the SEM structure simpler, probably by eliminating non-significant paths, reducing multicollinearity, or more thoroughly investigating latent constructs. Other SEM techniques or outside clinical data to enhance model generalizability and accuracy.

5 References

[1] Härdle, Wolfgang and Simar, Léopold (2007). *Applied multivariate statistical analysis.* Springer Berlin.

Factor Analysis: https://www.geeksforgeeks.org/factor-analysis-in-r-programming/
Discriminant Analysis : https://rpubs.com/Nolan/298913
Canonical Correlation Analysis : https://stats.oarc.ucla.edu/r/dae/canonical-correlation-analysis/
Structural Equation Modelling : https://www.google.com/search?q=structural+equation+modeling+r+code&rlz=1C1YTUH_enLK1096LK1096&oq=str&gs_lcrp=EgZjaHJvbWUqBggAEEUYOzIGCAAQRRg chrome&ie=UTF-8

6 Appendices

- **Dataset Link**
  https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic
- **R Coding**

```r
#Load the dataset
breast_cancer<- read.table("D:/4th Year/Semester 01/DSC4033
    Multivariate Methods II/Tutorial/Mini Project/wdbc.data", header
    = FALSE, sep=",")

#Check the NA value
```

```r
colSums(is.na(breast_cancer))
```

Principal Component Analysis

```r
#PCA
#remove the unwanted column
breast_cancer<- breast_cancer[,3:32]
#Check the correlation between varibles
cor_matrix<- cor(breast_cancer, method = "spearman")
ggcorrplot(cor_matrix)
#Remove the variable which has highest correlation with others
breast_cancer <- subset(breast_cancer,select = -c(1,11,21))
#Standardize the data
scaled_breast_cancer<- apply(breast_cancer, 2, scale)
#Calculate the eigen value and eigen vector
breast_cancer_cov<- cov(scaled_breast_cancer)
breast_cancer_eigen<- eigen(breast_cancer_cov)
breast_cancer_eigen
#Bulit the PCA function
pca_result <- prcomp(breast_cancer, scale. = TRUE)
#principal Components Score
pca_result$x
#variation eplained by each principal components
VE<-pca_result$sdev^2
PVE <- VE / sum(VE)
round(PVE, 3)
#Scree plot
fviz_eig(pca_result, addlabels = TRUE, ylim = c(0, 50))
#loading matrix
pca_result$rotation[,1:6]
```

Factor Analysis

```r
#FA
#Kaiser-Meyer-Oklin Test (KMO)
KMO(scaled_breast_cancer)
new_breast_cancer<- scaled_breast_cancer[,
   KMO(scaled_breast_cancer)$MSAi > 0.5]
#Bartkett's Test
cortest.bartlett(new_breast_cancer)
#Parallel analysis and scree plots
fa.parallel(breast_cancer,fm="pa",fa="fa")
#Factor Analysis using 'Principal Component Method'
breast_cancer_PC<- fa(breast_cancer_cov ,nfactors = 6,rotate =
   "none",n.obs = 500 ,covar = TRUE,fm = "pa")
breast_cancer_PC
#breastcancer_PC$loadings
as.data.frame(unclass(breast_cancer_PC$loadings))
#Communalities
as.data.frame(unclass(breast_cancer_PC$communality))
```

```
#Factor Analysis Using 'Maximum Likelihood Method
breast_cancer_ML <- fa(breast_cancer_cov,nfactors = 6,rotate =
   "none",n.obs = 500 , covar = TRUE, fm = 'ml')
breast_cancer_ML
#unrotated ML loading
as.data.frame(unclass(breast_cancer_ML$loadings))
#ML communalities
as.data.frame(unclass(breast_cancer_ML$communality))
#Rotate the PC loading using varimax method
fa(breast_cancer_cov ,nfactors = 6,rotate = "varimax",n.obs = 500
   ,covar = TRUE,fm = 'pa')
#Graph Factor Loading Matrices
fa.diagram(breast_cancer_PC_rotate)
```

Canonical Correlation Analysis

```
#CCA
X <- breast_cancer_data[,3:12]
Y <- breast_cancer_data[,13:22]
#canonical correlation
cc1 <- cc(X, Y)
cc1$cor
#canonical coefficient
cc1[3:4]
#Check the canonical dimention
rho <- cc1$cor
n <- dim(X)[1]
p <- length(X)
q<- length(Y)
p.asym(rho, n, p, q, tstat = "Wilks")
```

Discriminant Analysis

```
#DA
#Bartlett's Test
bartlett.test(radius1~Diagnosis)
bartlett.test(texture1~Diagnosis)
#QDA
training_sample <- sample(c(TRUE, FALSE), nrow(da_breast_cancer),
   replace = T, prob = c(0.8,0.2))
train <- da_breast_cancer[training_sample, ]
test <- da_breast_cancer[!training_sample,]
test <- da_breast_cancer[!training_sample, ]
#create model
qda_model <- qda(Diagnosis~ radius1 + texture1 + perimeter1  + area1
   + smoothness1  + compactness1 + concavity1 + concave_points1 +
   symmetry1 + fractal_dimension1 + radius2 + texture2 + perimeter2
   + area2 + smoothness2 + compactness2 + concavity2 +
   concave_points2 + symmetry2 +fractal_dimension2 + radius3 +
   texture3 + perimeter3  + area3 + smoothness3 + compactness3 +
```

```
    concavity3 + concave_points3 + symmetry3 + fractal_dimension3,
    data = train)
#qda prediction
qda.test <- predict(qda_model,test)
test$qda <- qda.test$class
table(test$qda,test$Diagnosis)
```

Structural Equation Modelling

```
#SEM
scale_sem_data <- sem_data %>%
  dplyr::select( radius1, texture1, perimeter1, area1, smoothness1,
     compactness1, concavity1, concave_points1, symmetry1,
     fractal_dimension1, radius2, texture2, perimeter2, area2,
     smoothness2, compactness2, concavity2, concave_points2,
     symmetry2, fractal_dimension2, radius3, texture3, perimeter3,
     area3, smoothness3, compactness3, concavity3, concave_points3,
     symmetry3, fractal_dimension3) %>%
  scale() %>% as.data.frame()
  sem_data<- cbind(Diagnosis, scale_sem_data)
#convert categorical variable as factor
sem_data$Diagnosis <- ifelse(sem_data$Diagnosis == "M", 1, 0)
#model
model_1<-  '
 compactness1 ~ perimeter1 + area1
 compactness2 ~  area2
 compactness3 ~  area3
 shape1 =~ radius1 + texture1  + smoothness1 + compactness1 +
    concavity1 + concave_points1 + symmetry1 + fractal_dimension1
 shape2 =~ radius2 + texture2 + smoothness2 + compactness2 +
    concavity2 + concave_points2 + symmetry2 + fractal_dimension2
 shape3 =~ radius3 + texture3  + smoothness3 + compactness3 +
    concavity3 + concave_points3 + symmetry3 + fractal_dimension3
 Diagnosis ~  shape1 + shape2 + shape3
 '
fit_1 <- sem(model_1, data=sem_data)
summary(fit_1, fit.measures=TRUE)
#paths
semPaths(fit_1, whatLabels = "std", layout = "tree", edge.label.cex
   = 0.8)
```