Multivariate Analysis of diabetes data

Jas Sur, Hassasn Sarr, Pushpalatha Solligi, Priyadarshini Thota

5/6/2022

Reducing the dimesionality of diabetes factors by using PCA, factor analysis and non-metric MDS techniques

Data Source

source : http : //odds.cs.stonybrook.edu/pima − indians − diabetes − dataset

Following data is observed and used for predicting if a person has diabetes.

```
diabetes <- readMat('C:/Users/JasSu/Downloads/pima.mat')
diabetesX <- data.frame(diabetes$X)
diabetesY <- data.frame(diabetes$y)
colnames(diabetesY) <- "Y"
colnames(diabetesX)<- c("Pregnancies", "Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI", "
str(diabetesX)
```

```
## 'data.frame':    768 obs. of  8 variables:
##  $ Pregnancies             : num  6 1 8 1 0 5 3 10 2 8 ...
##  $ Glucose                 : num  148 85 183 89 137 116 78 115 197 125 ...
##  $ BloodPressure           : num  72 66 64 66 40 74 50 0 70 96 ...
##  $ SkinThickness           : num  35 29 0 23 35 0 32 0 45 0 ...
##  $ Insulin                 : num  0 0 0 94 168 0 88 0 543 0 ...
##  $ BMI                     : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
##  $ DiabetesPedigreeFunction: num  0.627 0.351 0.672 0.167 2.288 ...
##  $ Age                     : num  50 31 32 21 33 30 26 29 53 54 ...
```

```
str(diabetesY$Y)
```

```
##  num [1:768] 1 0 1 0 1 0 1 0 1 1 ...
```

```
summary(diabetesX)
```

```
##   Pregnancies        Glucose       BloodPressure     SkinThickness
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
```

```
## 3rd Qu.: 6.000    3rd Qu.:140.2    3rd Qu.: 80.00    3rd Qu.:32.00
## Max.   :17.000    Max.   :199.0    Max.   :122.00    Max.   :99.00
##     Insulin            BMI       DiabetesPedigreeFunction      Age
## Min.   :  0.0    Min.   : 0.00    Min.   :0.0780          Min.   :21.00
## 1st Qu.:  0.0    1st Qu.:27.30    1st Qu.:0.2437          1st Qu.:24.00
## Median : 30.5    Median :32.00    Median :0.3725          Median :29.00
## Mean   : 79.8    Mean   :31.99    Mean   :0.4719          Mean   :33.24
## 3rd Qu.:127.2    3rd Qu.:36.60    3rd Qu.:0.6262          3rd Qu.:41.00
## Max.   :846.0    Max.   :67.10    Max.   :2.4200          Max.   :81.00
```

```
#m <- dist(diabetesX[,1:8])
#m1 <- as.matrix(m)[1:5,1:8]

#palette(rainbow(8)) # set colors
#stars( diabetesX, len=1, cex=0.5, key.loc=c(12.5, 2),
#labels=row.names(diabetesX), draw.segments=TRUE)
```

Check for the correlation between

```
pairs(diabetesX)
```
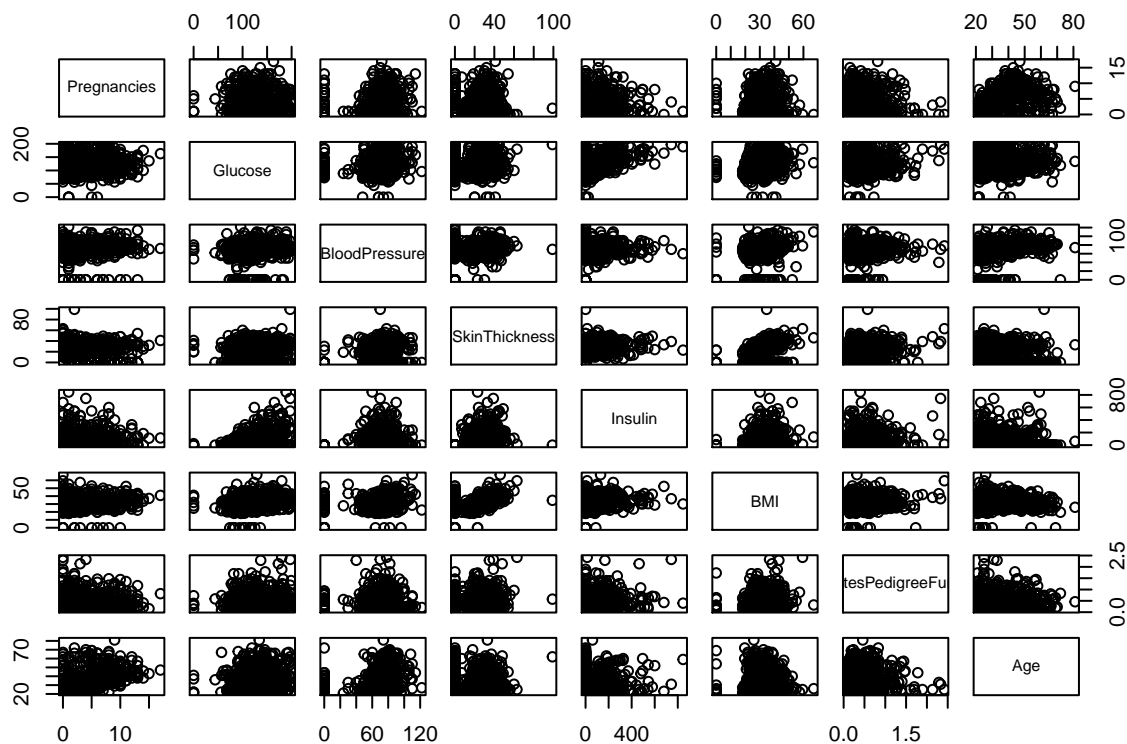
Reduce dependent variables which are correlated with each other.

```
diabetes.pc <- princomp(scale(diabetesX), cor = T)
summary(diabetes.pc)
```

```
## Importance of components:
##                            Comp.1    Comp.2    Comp.3    Comp.4      Comp.5
## Standard deviation     1.4471973 1.3157546 1.0147068 0.9356971 0.87312335
## Proportion of Variance 0.2617975 0.2164013 0.1287037 0.1094411 0.09529305
## Cumulative Proportion  0.2617975 0.4781988 0.6069025 0.7163436 0.81163667
##                            Comp.6     Comp.7     Comp.8
## Standard deviation     0.82621328 0.64793223 0.63597331
## Proportion of Variance 0.08532855 0.05247702 0.05055776
## Cumulative Proportion  0.89696522 0.94944224 1.00000000
```

```
sapply(diabetesX, sd)
```

```
##                Pregnancies                      Glucose          BloodPressure
##                  3.3695781                   31.9726182             19.3558072
##              SkinThickness                      Insulin                    BMI
##                 15.9522176                  115.2440024              7.8841603
## DiabetesPedigreeFunction                          Age
##                  0.3313286                   11.7602315
```

```
loadings<- diabetes.pc$loadings
print(loadings)
```

```
##
## Loadings:
##                          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## Pregnancies               0.128  0.594                0.476  0.194  0.589
## Glucose                   0.393  0.174 -0.468 -0.404 -0.466
## BloodPressure             0.360  0.184  0.535        -0.328 -0.634  0.192
## SkinThickness             0.440 -0.332  0.238         0.488        -0.282
## Insulin                   0.435 -0.251 -0.337 -0.350  0.347 -0.271  0.132
## BMI                       0.452 -0.101  0.362        -0.253  0.685
## DiabetesPedigreeFunction  0.271 -0.122 -0.433  0.834 -0.120
## Age                       0.198  0.621                0.109        -0.712
##                          Comp.8
## Pregnancies               0.118
## Glucose                   0.450
## BloodPressure
## SkinThickness             0.566
## Insulin                  -0.549
## BMI                      -0.342
## DiabetesPedigreeFunction
## Age                      -0.212
##
##                  Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## SS loadings       1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var    0.125  0.125  0.125  0.125  0.125  0.125  0.125  0.125
## Cumulative Var    0.125  0.250  0.375  0.500  0.625  0.750  0.875  1.000
```

```r
print(loadings, digits=2, cutoff=0.2, sort=T)
```

```
##
## Loadings:
##                         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## Pregnancies                     0.59                 0.48          0.59
## DiabetesPedigreeFunction  0.27        -0.43   0.83
## BloodPressure             0.36         0.54        -0.33  -0.63
## BMI                       0.45         0.36        -0.25   0.69
## Age                              0.62                            -0.71
## SkinThickness             0.44  -0.33   0.24         0.49        -0.28
## Insulin                   0.44  -0.25  -0.34  -0.35   0.35  -0.27
## Glucose                   0.39        -0.47  -0.40  -0.47
##                         Comp.8
## Pregnancies
## DiabetesPedigreeFunction
## BloodPressure
## BMI                      -0.34
## Age                      -0.21
## SkinThickness             0.57
## Insulin                  -0.55
## Glucose                   0.45
##
##                 Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## SS loadings       1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00
## Proportion Var    0.12   0.13   0.12   0.12   0.13   0.12   0.13   0.12
## Cumulative Var    0.12   0.25   0.37   0.50   0.62   0.75   0.88   1.00
```
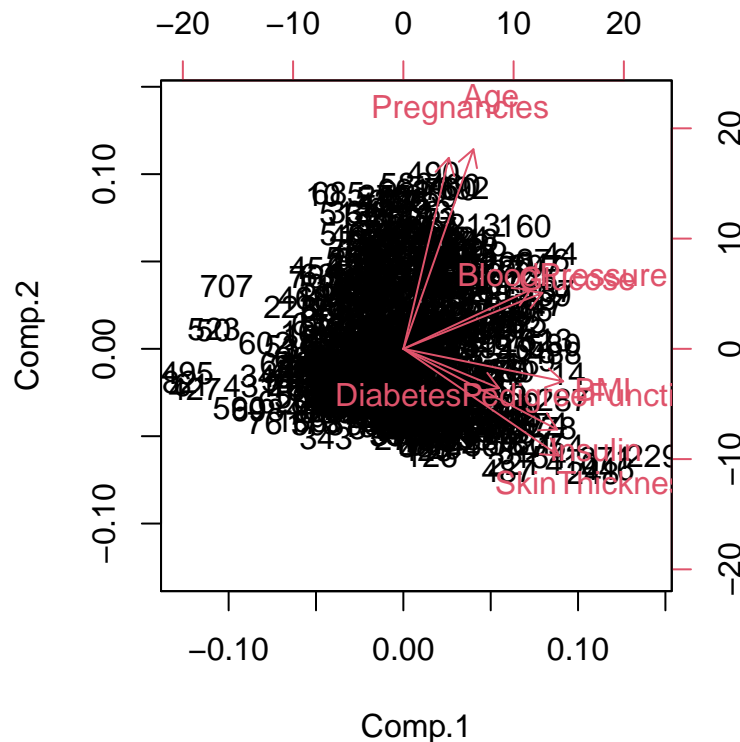
```r
screeplot(diabetes.pc, col = "red", pch = 16,
          type = "lines", cex = 2, lwd = 2, main = "")
abline(h=1, col='blue', lwd=1, lty=5)
```
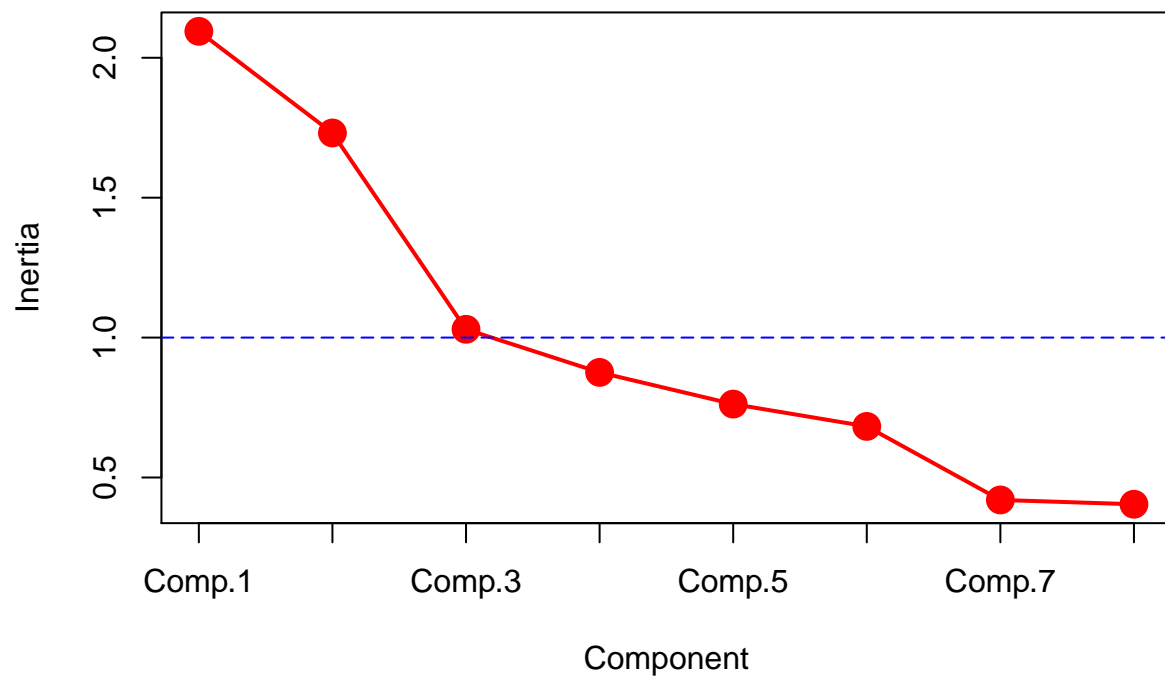
```
biplot(diabetes.pc)
```

ScreePlot suggest that 3 should be sufficient to represent the model with var>=1, Looking at the biplit, it can be ascertained that Pregencies and Age contribute to comp2, and factors other than Blood Pressure Glucose are orthogonal to it.

We can check these factors further by doing Factor analysis.

Exploitary check using Factor analysis

```
diabetes.std = as.data.frame(diabetesX)
str(diabetes.std)
```

```
## 'data.frame':    768 obs. of  8 variables:
##  $ Pregnancies             : num  6 1 8 1 0 5 3 10 2 8 ...
##  $ Glucose                 : num  148 85 183 89 137 116 78 115 197 125 ...
##  $ BloodPressure           : num  72 66 64 66 40 74 50 0 70 96 ...
##  $ SkinThickness           : num  35 29 0 23 35 0 32 0 45 0 ...
##  $ Insulin                 : num  0 0 0 94 168 0 88 0 543 0 ...
##  $ BMI                     : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
##  $ DiabetesPedigreeFunction: num  0.627 0.351 0.672 0.167 2.288 ...
##  $ Age                     : num  50 31 32 21 33 30 26 29 53 54 ...
```

```
fact = factanal(diabetes.std, factors=3, rotation = 'none', na.omit=TRUE)
fact
```

```
## 
## Call:
## factanal(x = diabetes.std, factors = 3, rotation = "none", na.omit = TRUE)
## 
## Uniquenesses:
##             Pregnancies                   Glucose            BloodPressure
##                   0.614                     0.005                    0.845
##           SkinThickness                   Insulin                      BMI
##                   0.188                     0.674                    0.767
## DiabetesPedigreeFunction                      Age
##                   0.939                     0.230
## 
## Loadings:
##                          Factor1 Factor2 Factor3
## Pregnancies               -0.298   0.133   0.529
## Glucose                            0.997
## BloodPressure              0.125   0.155   0.339
## SkinThickness              0.849           0.297
## Insulin                    0.459   0.333
## BMI                        0.379   0.223   0.199
## DiabetesPedigreeFunction   0.187   0.138
## Age                       -0.408   0.268   0.729
## 
##                Factor1 Factor2 Factor3
## SS loadings      1.381   1.292   1.065
## Proportion Var   0.173   0.161   0.133
## Cumulative Var   0.173   0.334   0.467
## 
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 39 on 7 degrees of freedom.
## The p-value is 1.95e-06
```

```
#three factors are not sufficnet
fact = factanal(diabetes.std, factors=4, rotation = 'none', na.omit=TRUE)
fact
```

```
## 
## Call:
## factanal(x = diabetes.std, factors = 4, rotation = "none", na.omit = TRUE)
## 
## Uniquenesses:
##             Pregnancies                   Glucose            BloodPressure
##                   0.658                     0.592                    0.813
##           SkinThickness                   Insulin                      BMI
##                   0.005                     0.472                    0.356
## DiabetesPedigreeFunction                      Age
##                   0.924                     0.119
## 
## Loadings:
##                          Factor1 Factor2 Factor3 Factor4
## Pregnancies                        0.570
## Glucose                            0.321   0.439   0.330
## BloodPressure              0.208   0.297   0.191  -0.139
## SkinThickness              0.997
```

```
## Insulin                       0.440          0.301   0.494
## BMI                           0.396   0.131   0.602  -0.327
## DiabetesPedigreeFunction  0.185          0.154   0.119
## Age                          -0.118   0.929
##
##               Factor1 Factor2 Factor3 Factor4
## SS loadings     1.447   1.402   0.716   0.497
## Proportion Var  0.181   0.175   0.090   0.062
## Cumulative Var  0.181   0.356   0.446   0.508
##
## Test of the hypothesis that 4 factors are sufficient.
## The chi square statistic is 2.9 on 2 degrees of freedom.
## The p-value is 0.235
```

```
print(fact, digits=2, cutoff=0.3, sort=T)
```

```
##
## Call:
## factanal(x = diabetes.std, factors = 4, rotation = "none", na.omit = TRUE)
##
## Uniquenesses:
##           Pregnancies                Glucose        BloodPressure
##                  0.66                   0.59                 0.81
##         SkinThickness                Insulin                  BMI
##                  0.00                   0.47                 0.36
## DiabetesPedigreeFunction                Age
##                  0.92                   0.12
##
## Loadings:
##                    Factor1 Factor2 Factor3 Factor4
## SkinThickness         1.00
## Pregnancies                   0.57
## Age                           0.93
## BMI                   0.40            0.60   -0.33
## Glucose                       0.32    0.44    0.33
## BloodPressure
## Insulin               0.44            0.30    0.49
## DiabetesPedigreeFunction
##
##               Factor1 Factor2 Factor3 Factor4
## SS loadings      1.45    1.40    0.72    0.50
## Proportion Var   0.18    0.18    0.09    0.06
## Cumulative Var   0.18    0.36    0.45    0.51
##
## Test of the hypothesis that 4 factors are sufficient.
## The chi square statistic is 2.9 on 2 degrees of freedom.
## The p-value is 0.235
```

It looks like factors 3 are insufficient and factor 4 are sufficient to explain the model. We fail to reject the null hypothesis that the factor 4 would be sufficient to represent a reduced dimensions model

Check residuals to be small so that the reduced dimensions are sufficient to represent a reduced model.

Rotate the model

```
fact = factanal(diabetes.std, factors=4, rotation = 'varimax', na.omit=TRUE)
fact
```

```
##
## Call:
## factanal(x = diabetes.std, factors = 4, rotation = "varimax",    na.omit = TRUE)
##
## Uniquenesses:
##            Pregnancies                 Glucose          BloodPressure
##                  0.658                   0.592                  0.813
##          SkinThickness                 Insulin                    BMI
##                  0.005                   0.472                  0.356
## DiabetesPedigreeFunction                   Age
##                  0.924                   0.119
##
## Loadings:
##                          Factor1 Factor2 Factor3 Factor4
## Pregnancies               0.582
## Glucose                   0.205  -0.115   0.569   0.171
## BloodPressure             0.236   0.125           0.325
## SkinThickness                     0.922   0.238   0.280
## Insulin                           0.282   0.662
## BMI                               0.141   0.195   0.766
## DiabetesPedigreeFunction          0.110   0.233
## Age                       0.929
##
##                  Factor1 Factor2 Factor3 Factor4
## SS loadings        1.319   0.997   0.931   0.816
## Proportion Var     0.165   0.125   0.116   0.102
## Cumulative Var     0.165   0.289   0.406   0.508
##
## Test of the hypothesis that 4 factors are sufficient.
## The chi square statistic is 2.9 on 2 degrees of freedom.
## The p-value is 0.235
```

```
print(fact, digits=2, cutoff=0.3, sort=T)
```

```
##
## Call:
## factanal(x = diabetes.std, factors = 4, rotation = "varimax",    na.omit = TRUE)
##
## Uniquenesses:
##            Pregnancies                 Glucose          BloodPressure
##                   0.66                    0.59                   0.81
##          SkinThickness                 Insulin                    BMI
##                   0.00                    0.47                   0.36
## DiabetesPedigreeFunction                   Age
##                   0.92                    0.12
##
## Loadings:
```

```
##                       Factor1 Factor2 Factor3 Factor4
## Pregnancies            0.58
## Age                    0.93
## SkinThickness                  0.92
## Glucose                                0.57
## Insulin                                0.66
## BMI                                            0.77
## BloodPressure                                  0.33
## DiabetesPedigreeFunction
##
##              Factor1 Factor2 Factor3 Factor4
## SS loadings    1.32    1.00    0.93    0.82
## Proportion Var 0.16    0.12    0.12    0.10
## Cumulative Var 0.16    0.29    0.41    0.51
##
## Test of the hypothesis that 4 factors are sufficient.
## The chi square statistic is 2.9 on 2 degrees of freedom.
## The p-value is 0.235
```

It's much clear that the Factor1 can be explained by Pregnancies and Age, Factor2 by Skinthickness, Factor 3 by Flucose and Insulin, and Factor4 can be explained by BloodPressure and DiabetesPedigreeFunction.

The test provides a p-value of 0.235, which suggests that a 4 dimensions reduction of the data is sufficient at 5% significance level.

Check residuals

```
L = fact$loadings; L;  Psi = diag(fact$uniquenesses);Psi
```

```
##
## Loadings:
##                       Factor1 Factor2 Factor3 Factor4
## Pregnancies            0.582
## Glucose                0.205  -0.115   0.569   0.171
## BloodPressure          0.236   0.125           0.325
## SkinThickness                  0.922   0.238   0.280
## Insulin                        0.282   0.662
## BMI                            0.141   0.195   0.766
## DiabetesPedigreeFunction       0.110   0.233
## Age                    0.929
##
##              Factor1 Factor2 Factor3 Factor4
## SS loadings    1.319   0.997   0.931   0.816
## Proportion Var 0.165   0.125   0.116   0.102
## Cumulative Var 0.165   0.289   0.406   0.508
```

```
##          [,1]     [,2]      [,3]  [,4]      [,5]      [,6]      [,7]      [,8]
## [1,] 0.6578437 0.000000 0.0000000 0.000 0.0000000 0.0000000 0.0000000 0.000000
## [2,] 0.0000000 0.591578 0.0000000 0.000 0.0000000 0.0000000 0.0000000 0.000000
## [3,] 0.0000000 0.000000 0.8127369 0.000 0.0000000 0.0000000 0.0000000 0.000000
## [4,] 0.0000000 0.000000 0.0000000 0.005 0.0000000 0.0000000 0.0000000 0.000000
## [5,] 0.0000000 0.000000 0.0000000 0.000 0.4716588 0.0000000 0.0000000 0.000000
```

```
## [6,] 0.0000000 0.000000 0.0000000 0.000 0.0000000 0.3558406 0.0000000 0.000000
## [7,] 0.0000000 0.000000 0.0000000 0.000 0.0000000 0.0000000 0.9237768 0.000000
## [8,] 0.0000000 0.000000 0.0000000 0.000 0.0000000 0.0000000 0.0000000 0.119118
```

```r
resids = cor(diabetesX) - (L%*%t(L) + Psi)  ; resids
```

```
##                              Pregnancies        Glucose BloodPressure
## Pregnancies                 1.318673e-06  5.994472e-03 -4.994879e-03
## Glucose                     5.994472e-03 -9.223485e-07  6.603275e-03
## BloodPressure              -4.994879e-03  6.603275e-03  7.026252e-08
## SkinThickness               2.702657e-05 -6.497846e-06  2.705642e-05
## Insulin                     1.271592e-03  8.830411e-04 -4.408838e-04
## BMI                         1.088344e-03 -1.027914e-03 -5.630390e-05
## DiabetesPedigreeFunction   -3.516193e-02 -1.079072e-03 -2.928254e-02
## Age                         6.612853e-05 -9.240893e-04  3.603548e-04
##                            SkinThickness       Insulin           BMI
## Pregnancies                 2.702657e-05  1.271592e-03  1.088344e-03
## Glucose                    -6.497846e-06  8.830411e-04 -1.027914e-03
## BloodPressure               2.705642e-05 -4.408838e-04 -5.630390e-05
## SkinThickness              -5.265443e-08  3.221339e-06 -4.305107e-06
## Insulin                     3.221339e-06 -5.297747e-09  2.999977e-05
## BMI                        -4.305107e-06  2.999977e-05  7.209581e-08
## DiabetesPedigreeFunction    2.816875e-05 -3.204163e-03  4.977742e-03
## Age                        -3.888883e-06 -1.552557e-04 -9.100591e-05
##                            DiabetesPedigreeFunction          Age
## Pregnancies                            -3.516193e-02  6.612853e-05
## Glucose                                -1.079072e-03 -9.240893e-04
## BloodPressure                          -2.928254e-02  3.603548e-04
## SkinThickness                           2.816875e-05 -3.888883e-06
## Insulin                                -3.204163e-03 -1.552557e-04
## BMI                                     4.977742e-03 -9.100591e-05
## DiabetesPedigreeFunction               -1.469487e-06  5.143935e-03
## Age                                     5.143935e-03  1.220708e-07
```

```r
summary(resids[row(resids)!=col(resids)])
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -3.516e-02 -5.617e-04 -3.300e-07 -1.783e-03  4.910e-04  6.603e-03
```

The residuals are small so we can ascertain that the reduced dimension represent original model fairly well with around 51 of variance.

Verify factors with non-meteric MDS

Since this is non-distance (metric) data, we checked goodness of fit using Shepard and Stress plots.

```r
D <- dist(diabetesX)
head(D)
```

```
## [1]  66.90348  54.29634 115.73012 172.58913  52.13618 117.17040
```

```
#
#
# mds <- isoMDS(D)
#
# stress2 = vector(length = 7)
# for(k in 1:length(stress2)) {
#     fit2 = isoMDS(D, y = cmdscale(D, k), k, trace = FALSE)
#     stress2[k] = fit2$stress
# }
# plot(stress2, main = "Stress vs. k") # k = 2 is good enough?
#
# x <- mds$points[,1]
# y <- mds$points[,2]
#
# plot(x, y, xlab = "Coordinate 1", ylab = "Coordinate 2",  xlim = range(mds$points[,1])*1.2, type = "n
# text(x, y, labels = colnames(diabetesX), cex = 0.6)
#
# diabetes_sh <- Shepard(diabetesX[lower.tri(diabetesX)], mds$points)
#
# #x-axis=  orig proximity (similarity or dissimilarity) x(i,j)
# #y-axis = d(i,j) =asterisk
# plot(diabetes_sh, pch = "*", cex=1, col="red",
#     xlab = "Dissimilarity",
#     ylab = "Distance",
#     xlim = range(diabetes_sh$x),
#     ylim = range(diabetes_sh$x))
#
# #x-axis= orig proximity(similarity or dissimilarity)
# #y-axis= the fit( \hat{d(i,j)}) =dash
# lines(diabetes_sh$x, diabetes_sh$yf, type = "S", lwd=2,col="blue")
```

```
library(vegan)
diabetes_mds2  <- metaMDS(diabetesX, k=4)
```

```
## Square root transformation
## Wisconsin double standardization
## Run 0 stress 0.06121611
## Run 1 stress 0.06177514
## Run 2 stress 0.06122449
## ... Procrustes: rmse 0.0009288842  max resid 0.01537085
## Run 3 stress 0.0612629
## ... Procrustes: rmse 0.004047671  max resid 0.06568366
## Run 4 stress 0.06122993
## ... Procrustes: rmse 0.0006380424  max resid 0.01288919
## Run 5 stress 0.06123394
## ... Procrustes: rmse 0.0008310833  max resid 0.01344095
## Run 6 stress 0.06127028
## ... Procrustes: rmse 0.004125378  max resid 0.06583423
## Run 7 stress 0.06173031
## Run 8 stress 0.06125167
## ... Procrustes: rmse 0.003827134  max resid 0.06549385
## Run 9 stress 0.06122064
## ... Procrustes: rmse 0.0002877605  max resid 0.005498155
```

```
## ... Similar to previous best
## Run 10 stress 0.0621178
## Run 11 stress 0.06129146
## ... Procrustes: rmse 0.003438192  max resid 0.0651043
## Run 12 stress 0.06121638
## ... Procrustes: rmse 0.0007245537  max resid 0.008143436
## ... Similar to previous best
## Run 13 stress 0.06136674
## ... Procrustes: rmse 0.004540835  max resid 0.06631846
## Run 14 stress 0.06122968
## ... Procrustes: rmse 0.0008300005  max resid 0.01930606
## Run 15 stress 0.06129681
## ... Procrustes: rmse 0.004245805  max resid 0.06578371
## Run 16 stress 0.06124696
## ... Procrustes: rmse 0.001511969  max resid 0.01839689
## Run 17 stress 0.06137857
## ... Procrustes: rmse 0.00512532  max resid 0.06656211
## Run 18 stress 0.06126185
## ... Procrustes: rmse 0.003701733  max resid 0.06546292
## Run 19 stress 0.06125783
## ... Procrustes: rmse 0.003828786  max resid 0.06505558
## Run 20 stress 0.06122108
## ... Procrustes: rmse 0.0008207116  max resid 0.009888921
## ... Similar to previous best
## *** Solution reached
```
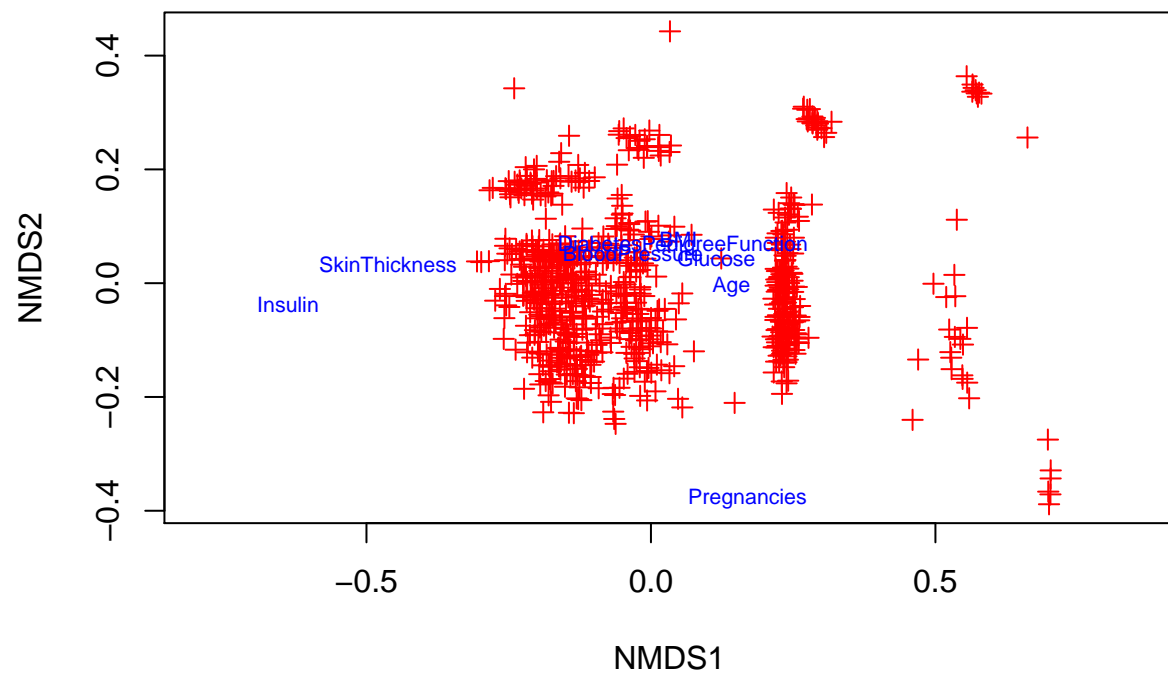
```
plot(diabetes_mds2, type="n")
points(diabetes_mds2, display=c("sites"), choices=c(1,2), pch=3, col="red")
text(diabetes_mds2, display=c("species"), choicec=c(1,2), col="blue", cex=0.7)
```
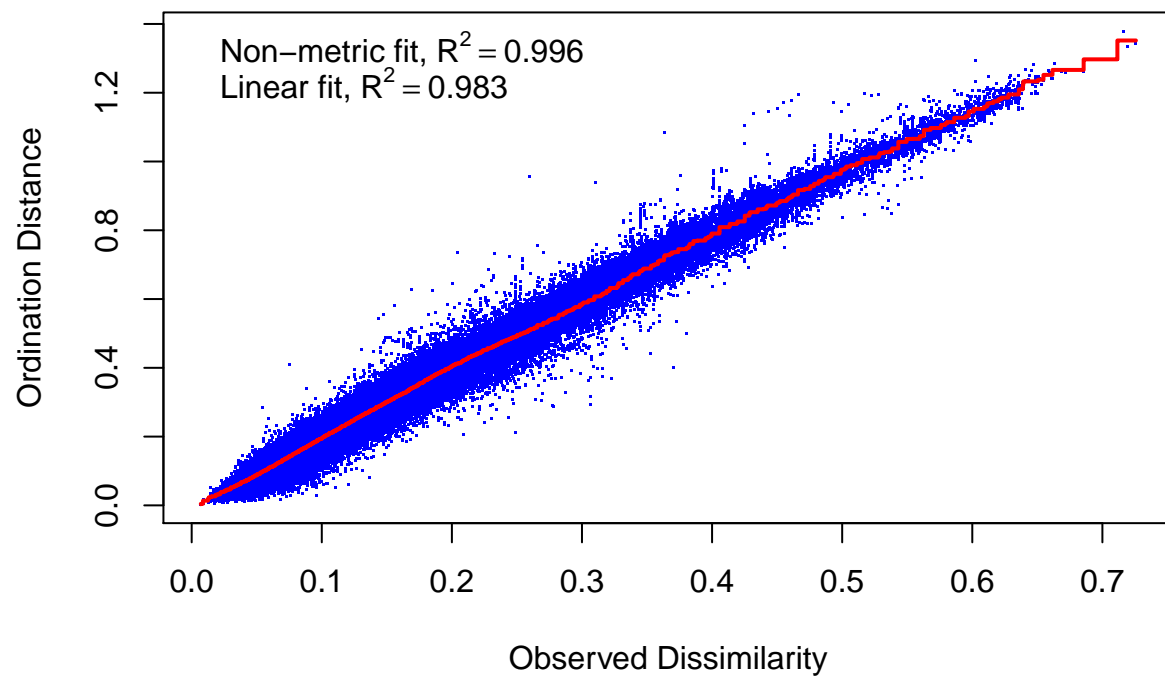
```
## Warning in text.default(x, labels = rownames(x), ...): "choicec" is not a
## graphical parameter
```

```
stressplot(diabetes_mds2)
```

Non−metric fit, $R^2 = 0.996$
Linear fit, $R^2 = 0.983$

The plot shows a R^2 value of 0.966, which is close to zero, so we can claim that there is a good fit.