Cervical Cancer

Richard Olekanma

12/29/2020

Cervical Cancer Analysis

This study attempts to use PCA and CFA to determine underlying Factors and meanings from our dataset. I also took this a step further and projected the PCA on to our old dataset to create a new "Cervical" dataset with lower dimensionality. One of the first things that needed to be done was preprocessing of our data frame. It was imperative to reduce noise in the data frame and we started by removing columns that were all 0's such as, location cervical condylomtosis was removed due to not having any impact on what we are trying to model as well as producing error messages with different R programs. Repetitive and redundant columns such as smoking (packs/year), STD time since first diagnosis, Time since last Diagnosis, STDs Number of Diagnosis and Hormonal contraceptives were also removed. These variables were highly correlated with each other and by removing them we eliminated a significant amount of multicollinearity in the dataset. These variables were also redundant as well because several of these columns were repeated in different columns with the same meaning. Various testing measures such as Hinselmann, Schiller, Citology, Biopsy, are typically recommended by a doctor when they think the patient has cervical cancer and the test is done to confirm. Our group decided that those columns were best aggregated into one column that indicated whether the patient had this time of test done or not.

##Data

```
## # A tibble: 6 x 28
##     Age `Number of sexu~ `First sexual i~ `Num of pregnan~ Smokes
##   <dbl>            <dbl>            <dbl>            <dbl>  <dbl>
## 1    18                4               15                1      0
## 2    15                1               14                1      0
## 3    34                1               18.2              1      0
## 4    52                5               16                4      1
## 5    46                3               21                4      0
## 6    42                3               23                2      0
## # ... with 23 more variables: `Smokes (years)` <dbl>, `Smokes
## #   (packs/year)` <dbl>, `Hormonal Contraceptives` <dbl>, `Hormonal
## #   Contraceptives (years)` <dbl>, IUD <dbl>, `IUD (years)` <dbl>,
## #   STDs <dbl>, `STDs (number)` <dbl>, `Location_vaginal
## #   condylomatosis` <dbl>, `Location_vulvo-perineal condylomatosis` <dbl>,
## #   `STDs: Number of diagnosis` <dbl>, `STDs: Time since first
## #   diagnosis` <dbl>, `STDs: Time since last diagnosis` <dbl>,
## #   has_cancer <dbl>, Abnormal_cell_growth_in_cervix <dbl>, has_HPV <dbl>,
## #   Cervical_cancer <dbl>, Hinselmann <dbl>, Schiller <dbl>,
## #   Citology <dbl>, Biopsy <dbl>, Number_of_tests <dbl>, Risk_exists <dbl>
```

Principal Component Analysis

PCA projects high dimensional data on to a lower space to reduce the dimensionality with eigenvalues and eigenvectors. This will reduce the complexitiy of a dataset and capture the maximum amount of

variance. Parsimony is key to interpretation, with a small number of components we can expect a small # of contributions to each.

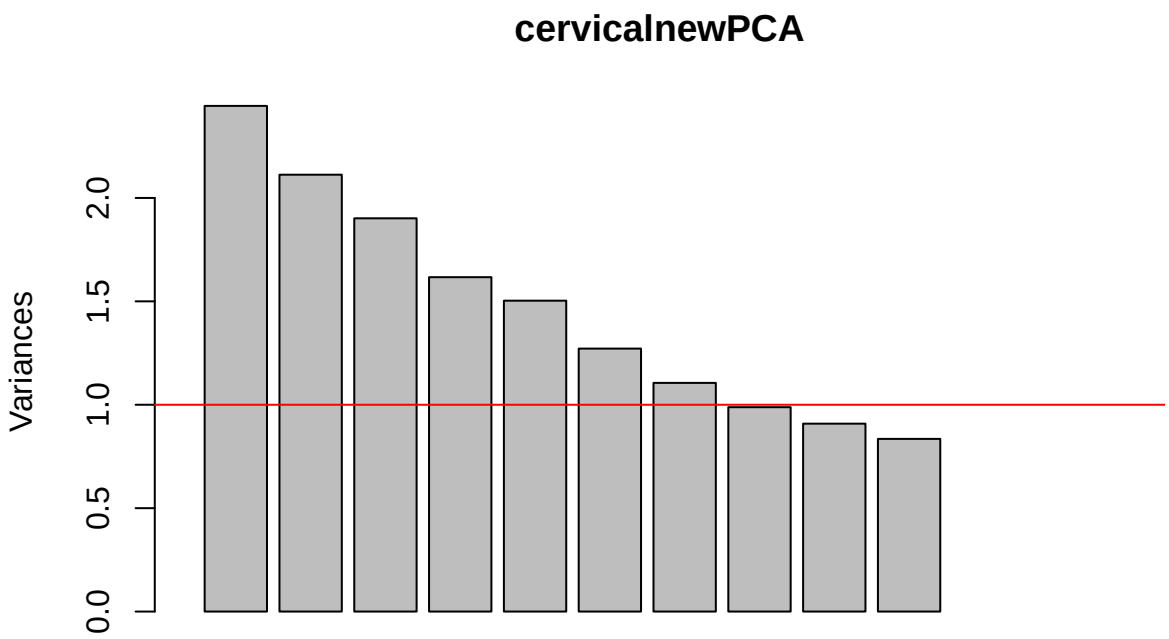
Running PCA on scaled Data

```
round(cervicalnewPCA$rotation, 2)
```

```
##                                      PC1   PC2   PC3   PC4   PC5   PC6
## Age                                 0.31 -0.20 -0.36 -0.24  0.13 -0.34
## Number of sexual partners           0.10 -0.02 -0.17  0.21 -0.21  0.09
## First sexual intercourse            0.06 -0.02 -0.07 -0.21  0.38 -0.29
## Num of pregnancies                  0.25 -0.18 -0.35 -0.19 -0.07 -0.18
## Smokes (years)                      0.23 -0.05 -0.30  0.41 -0.27  0.00
## Smokes (packs/year)                 0.23 -0.05 -0.27  0.44 -0.22  0.05
## Hormonal Contraceptives             0.12 -0.15 -0.13 -0.32 -0.01  0.58
## Hormonal Contraceptives (years)     0.23 -0.14 -0.18 -0.38 -0.04  0.43
## IUD (years)                         0.14 -0.04 -0.10 -0.06  0.04 -0.38
## STDs (number)                       0.11  0.60 -0.22 -0.01  0.12  0.09
## Location_vaginal condylomatosis     0.02  0.30 -0.15  0.03  0.21 -0.05
## Location_vulvo-perineal condylomatosis 0.10  0.58 -0.21 -0.04  0.11  0.11
## has_cancer                          0.42 -0.10  0.26  0.24  0.40  0.10
## Abnormal_cell_growth_in_cervix      0.03 -0.02  0.01 -0.10 -0.12 -0.13
## has_HPV                             0.42 -0.10  0.26  0.24  0.39  0.14
## Number_of_tests                     0.37  0.20  0.33 -0.19 -0.36 -0.12
## Risk_exists                         0.35  0.19  0.36 -0.19 -0.37 -0.12
##                                      PC7   PC8   PC9  PC10  PC11  PC12
## Age                                -0.06  0.04 -0.17  0.03 -0.08  0.11
## Number of sexual partners           0.33  0.07 -0.73  0.14  0.44 -0.05
## First sexual intercourse           -0.56  0.12 -0.28 -0.26  0.23 -0.13
## Num of pregnancies                  0.25 -0.04  0.02  0.33 -0.46 -0.27
## Smokes (years)                     -0.25  0.01  0.17 -0.13 -0.09 -0.08
## Smokes (packs/year)                -0.29  0.05  0.16 -0.11  0.13  0.09
## Hormonal Contraceptives            -0.05  0.00  0.14 -0.13  0.22 -0.62
## Hormonal Contraceptives (years)    -0.04 -0.03  0.07  0.01  0.11  0.69
## IUD (years)                         0.43 -0.40  0.37 -0.31  0.49 -0.01
## STDs (number)                       0.10  0.00 -0.03 -0.12 -0.13  0.00
## Location_vaginal condylomatosis    -0.12  0.17  0.31  0.72  0.40 -0.03
## Location_vulvo-perineal condylomatosis  0.10 -0.04 -0.08 -0.26 -0.14  0.02
## has_cancer                          0.13  0.03  0.01 -0.01 -0.03  0.01
## Abnormal_cell_growth_in_cervix      0.28  0.88  0.20 -0.24  0.06  0.01
## has_HPV                             0.12  0.06 -0.01  0.03 -0.09  0.00
## Number_of_tests                    -0.14 -0.07 -0.03  0.07  0.02 -0.08
## Risk_exists                        -0.12 -0.01 -0.02  0.06  0.06 -0.02
##                                     PC13  PC14  PC15  PC16  PC17
## Age                                -0.03  0.70 -0.05 -0.01  0.00
## Number of sexual partners          -0.05 -0.08  0.01  0.00  0.00
## First sexual intercourse            0.00 -0.42  0.02  0.02 -0.03
## Num of pregnancies                  0.16 -0.47 -0.01 -0.01  0.03
## Smokes (years)                     -0.69 -0.08 -0.01  0.00  0.00
## Smokes (packs/year)                 0.69  0.01  0.00  0.00  0.01
## Hormonal Contraceptives             0.02  0.17 -0.03 -0.02 -0.01
## Hormonal Contraceptives (years)    -0.08 -0.23  0.04  0.00 -0.01
## IUD (years)                        -0.02 -0.11  0.01  0.06 -0.02
## STDs (number)                       0.03  0.01  0.06 -0.08 -0.71
## Location_vaginal condylomatosis    -0.08  0.02  0.00  0.01  0.10
```

```
## Location_vulvo-perineal condylomatosis  0.02  0.01 -0.08  0.07  0.68
## has_cancer                             -0.02 -0.03  0.02 -0.70  0.08
## Abnormal_cell_growth_in_cervix         -0.01 -0.05  0.05  0.01  0.01
## has_HPV                                -0.02  0.01 -0.02  0.70 -0.07
## Number_of_tests                         0.01  0.05  0.70  0.03  0.07
## Risk_exists                             0.00 -0.03 -0.70 -0.03 -0.06
```

```
plot(cervicalnewPCA, xlim = c(0,15))
abline(h=1, col = 'red')
```



Factors of 7 were chosen after running PCA. The scree plot indicates there is a knee around 8-9 PC's, with the rest reducing to noise that is not important for our future model. Variance captured from 7-9 is from 70.3 to 81.4%, 70% is a bit on the lower end for the ideal amount of variance captured but observing the groupings of loadings and seeing groups with individual factors when we used 8 or 9 factors led the group to believe that 7 is a good choice. After scaling the data to find PC's above 1 yielded 7 factors.

```
#principal factor analysis
pcerviccalnew = psych::principal(cervical_X, rotate = 'varimax',
                                 nfactors = 7, scores = TRUE)

#location vaginal condolymotis is in its own group
summary(pcerviccalnew)
```

```
##
## Factor analysis with Call: psych::principal(r = cervical_X, nfactors = 7, rotate = "varimax",
##     scores = TRUE)
##
## Test of the hypothesis that 7 factors are sufficient.
## The degrees of freedom for the model is 38  and the objective function was  1.66
## The number of observations was  766  with Chi Square =  1251.43  with prob <  6.3e-238
##
## The root mean square of the residuals (RMSA) is  0.07
```

```
print( pcerviccalnew$loadings, cutoff = .4, sort = T)
```

```
##
```

```
## Loadings:
##                                     RC2    RC1    RC5    RC4    RC3
## STDs (number)                        0.954
## Location_vaginal condylomatosis      0.511
## Location_vulvo-perineal condylomatosis  0.925
## has_cancer                                 0.963
## has_HPV                                    0.959
## Number_of_tests                                   0.953
## Risk_exists                                       0.959
## Smokes (years)                                           0.865
## Smokes (packs/year)                                      0.867
## Age                                                             0.758
## Num of pregnancies                                              0.733
## IUD (years)                                                     0.609
## Hormonal Contraceptives
## Hormonal Contraceptives (years)
## First sexual intercourse
## Number of sexual partners
## Abnormal_cell_growth_in_cervix
##                                     RC6    RC7
## STDs (number)
## Location_vaginal condylomatosis
## Location_vulvo-perineal condylomatosis
## has_cancer
## has_HPV
## Number_of_tests
## Risk_exists
## Smokes (years)
## Smokes (packs/year)
## Age                                        0.401
## Num of pregnancies
## IUD (years)
## Hormonal Contraceptives              0.836
## Hormonal Contraceptives (years)      0.818
## First sexual intercourse                0.864
## Number of sexual partners              -0.482
## Abnormal_cell_growth_in_cervix
##
##                  RC2   RC1   RC5   RC4   RC3   RC6   RC7
## SS loadings    2.043 1.896 1.891 1.692 1.629 1.538 1.267
## Proportion Var 0.120 0.112 0.111 0.100 0.096 0.090 0.075
## Cumulative Var 0.120 0.232 0.343 0.442 0.538 0.629 0.703
```

Factor Analysis: This analysis provides information on factors taht are not directly observed but have influence in the dataset (underlying factors) It also assumes a certain % of unexplainable variance in the dataset.

- PC1: has all the factors positively correlated with Age, number of pregnancies, smokes, hormonal contraceptives, has cancer, has HPV, risk exists and number of tests being the largest values.

- PC2: Consists of Age, Number of pregnancies, Hormonal contraceptives and Hormonal contraceptives (years) being inversely related with STDS (number), Location vaginal condylomatosis, number of tests, risk exists and Location vulvo perineal condylomatosis.

- PC3: Age, number of sexual partners, number of pregnancies, smokes, hormonal contraceptives, STDs, location vaginal + vulvo condylomatosis are inversely related with has cancer, abnormal cell growth, has

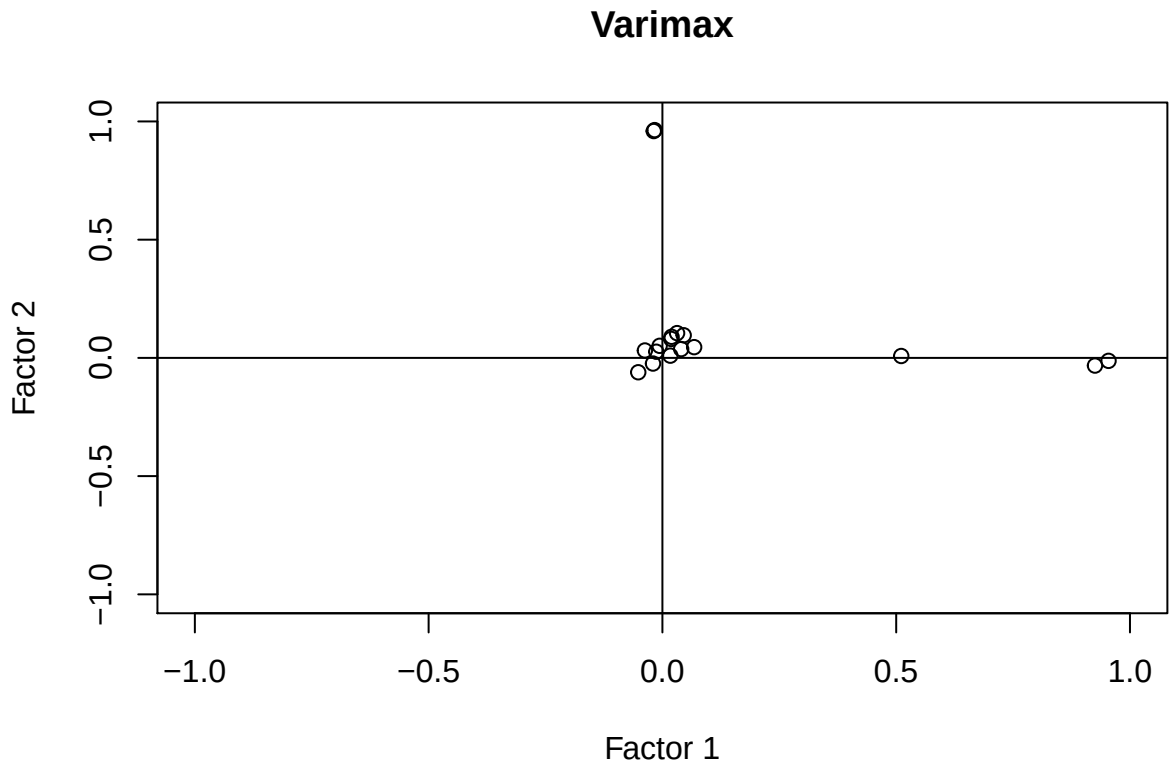
HPV, number of tests and Risk exists.

• PC4: Age, first sexual intercourse, number of pregnancies, hormonal contraceptives, Number of tests are negatively related with number of sexual partners, smoking, has cancer and has HPV.

• PC5: Consists of age, first sexual intercourse, location vaginal condylomatosis, has cancer, has HPV being inversely related with number of sexual partners, smoking, abnormal cell growth in cervix, number of tests and risk existing.

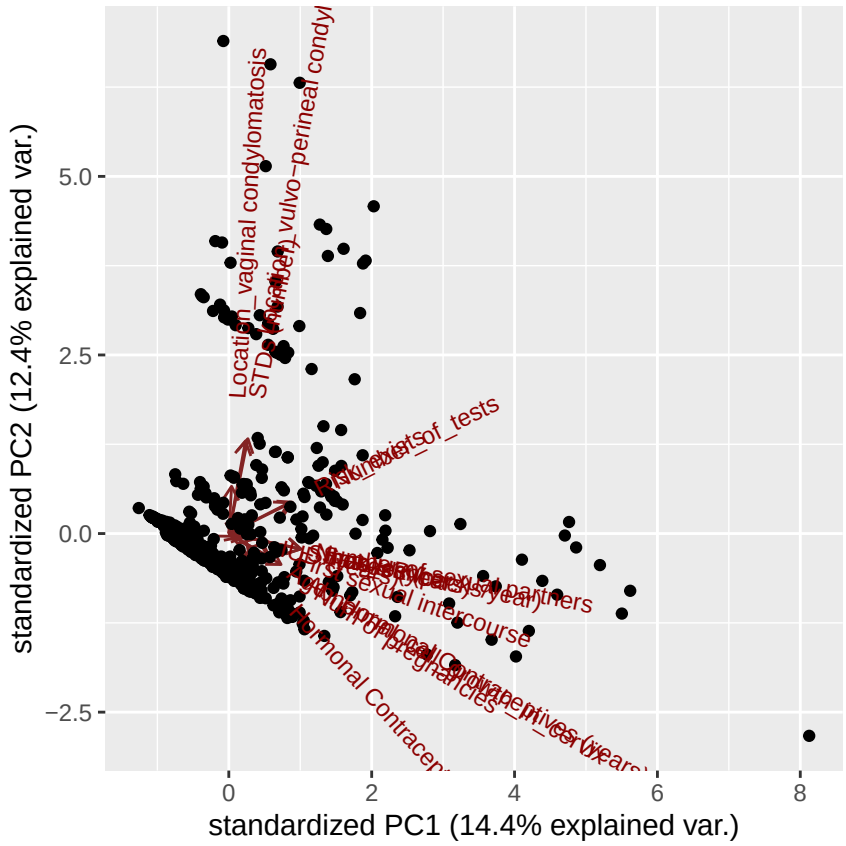
• PC6: Age, first sexual intercourse, number of pregnancies, IUD, abnormal cell growth, risk existing and number of tests are negatively related with hormonal contraceptives, has HPV and has cancer. • PC7: number of sexual partners, number of pregnancies, IUD, STDS, has cancer, abnormal cell growth, has HPV are inversely related with age, first sexual intercourse, smoking, location of vaginal condylomatosis.

```
plot(pcerviccalnew$loadings[,1],
     pcerviccalnew$loadings[,2],
     xlab = "Factor 1",
     ylab = "Factor 2",
     ylim = c(-1,1),
     xlim = c(-1,1),
     main = "Varimax")
abline(h = 0, v = 0)
```

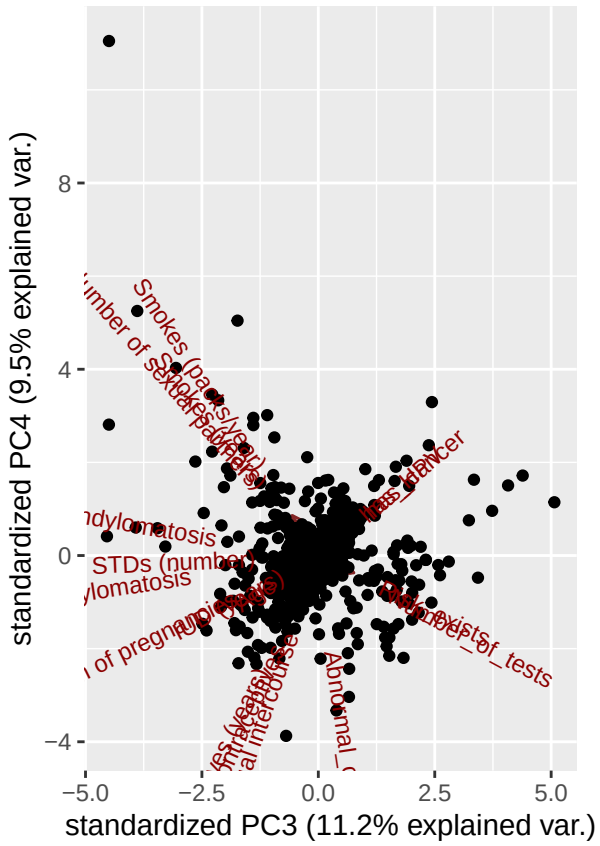


Using the Varimax algorithim that attempts to load variables highly or not at all on each factor. Can result in correlated factors but its main point is to redistrubute variance between factors.

```
ggbiplot(cervicalnewPCA,choices=c(1,2))
```




```
ggbiplot(cervicalnewPCA,choices=c(3,4))
```



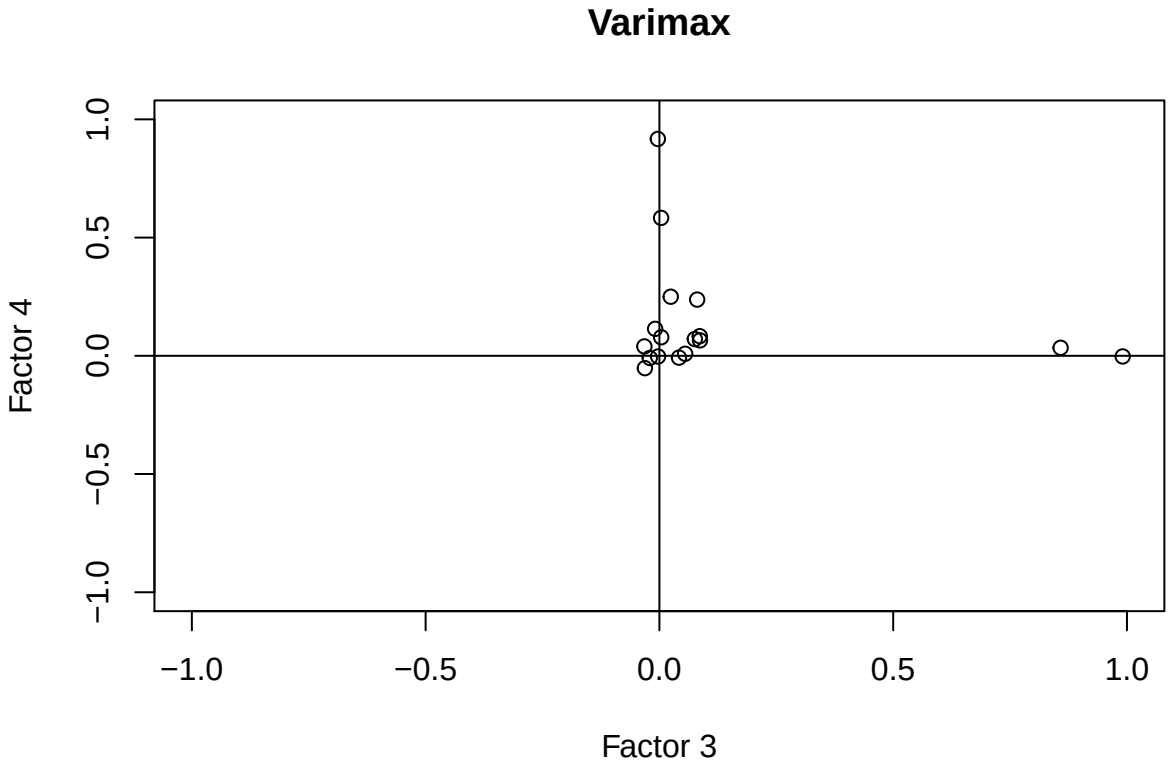
Common factor analysis on the other hand assumes unexplainable experimental error in the measurements within the emphasized shared variance. It uses the maximum likelyhood optimization, which specifies a specific number of factors to search for, changes factor coefficients to measure for improvements but this causes less speed and unique factors.

```
#common factor analysis
c = factanal(cervical_X, 8, rotation = "varimax")
print(c$loadings, cutoff = .4, sort = T)
```

```
##
## Loadings:
##                                          Factor1 Factor2 Factor3 Factor4
## STDs (number)                             0.946
## Location_vulvo-perineal condylomatosis    0.947
## has_cancer                                        0.990
## has_HPV                                           0.879
## Number_of_tests                                           0.858
## Risk_exists                                               0.991
## Age                                                                0.917
## Num of pregnancies                                                 0.583
## Smokes (years)
## Smokes (packs/year)
## First sexual intercourse
## Hormonal Contraceptives
## Hormonal Contraceptives (years)
## Number of sexual partners
## IUD (years)
## Location_vaginal condylomatosis
```

```
## Abnormal_cell_growth_in_cervix
##                                      Factor5 Factor6 Factor7 Factor8
## STDs (number)
## Location_vulvo-perineal condylomatosis
## has_cancer
## has_HPV
## Number_of_tests
## Risk_exists
## Age
## Num of pregnancies
## Smokes (years)                        0.717
## Smokes (packs/year)                   0.825
## First sexual intercourse                      0.988
## Hormonal Contraceptives                              0.868
## Hormonal Contraceptives (years)                      0.531
## Number of sexual partners
## IUD (years)
## Location_vaginal condylomatosis
## Abnormal_cell_growth_in_cervix
##
##                Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7
## SS loadings      1.863   1.792   1.753   1.341   1.294   1.124   1.097
## Proportion Var   0.110   0.105   0.103   0.079   0.076   0.066   0.065
## Cumulative Var   0.110   0.215   0.318   0.397   0.473   0.539   0.604
##                Factor8
## SS loadings      0.245
## Proportion Var   0.014
## Cumulative Var   0.618
```

```r
plot(c$loadings[,3],
     c$loadings[,4],
     xlab = "Factor 3",
     ylab = "Factor 4",
     ylim = c(-1,1),
     xlim = c(-1,1),
     main = "Varimax")
abline(h = 0, v = 0)
```



Common Factor Analysis on the 7 factors with varimax rotations, to try and get the original variables into distinct factors, by searching for the best n components near the eigenvectors with rotating the new axes the data will align better. Cutoff of .4 was set to clean up loadings and hide values with small loadings. The results yielded multiple loadings with only 1 factor having its own loading. CFA split exactly two into each loading with no cross contamination from the factors, each loading appears to be different lifestyle choices: using contraceptives, smoking, age and pregnancies. Condylomatosis both vulvo and vaginal grouped, indicating that if the person has one in the vulvo there may be one in the vaginal as well.

Classification

This analysis ends with classification of patients with or without cervical cancer.

```
new <- predict(pcerviccalnew, data = cervical_X)
head(new, 4)
```

```
##             RC2         RC1        RC5        RC4        RC3        RC6
## [1,] -0.2565007 -0.04133307 -0.3204273 -0.1377830 -0.6712526 -1.1111776
## [2,] -0.3056648 -0.07333542 -0.2203122 -0.3803266 -1.0469106 -1.1217142
## [3,] -0.2673861 -0.14219266 -0.2375851 -0.1579268 -0.1091173 -1.0947885
## [4,] -0.4065000  6.01206713 -0.7978792 14.4499464 -0.5934891  0.1058811
##             RC7
## [1,] -0.9257357
## [2,] -0.5645505
## [3,]  1.0788212
## [4,] -0.0232175
```

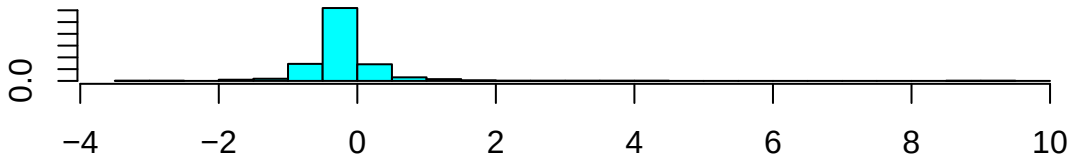
```
#adding cervical list to new pca matrix
newcervical <- cbind(new, cervical_Y)
```

Performing Linear Discriminate Analysis which tries to find a discriminate plane to evenly split the binary classes on.

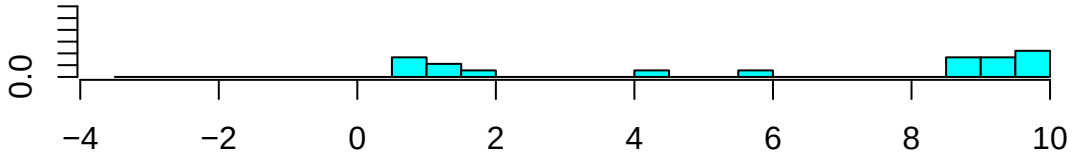
```r
#LDA on PCA matrix
PCALDA = lda(Cervical_cancer ~ ., data=newcervical)

#training test data
dt = sort(sample(nrow(newcervical), nrow(newcervical)*.8))
train<-newcervical[dt,]
test<-newcervical[-dt,]
```

```r
fit = lda(Cervical_cancer ~., data = train )#, CV = TRUE)
print(fit)
```

```
## Call:
## lda(Cervical_cancer ~ ., data = train)
##
## Prior probabilities of groups:
##          0          1
## 0.97058824 0.02941176
##
## Group means:
##           RC2        RC1         RC5         RC4         RC3          RC6
## 0 -0.03710954 -0.1174481 -0.02289365  0.04756793 -0.03293799  0.007175849
## 1 -0.26890822  3.8546258  0.66314855 -0.75213902  1.00260840 -0.215517810
##           RC7
## 0  0.01930226
## 1 -0.51419503
##
## Coefficients of linear discriminants:
##             LD1
## RC2 -0.09398926
## RC1  1.33722441
## RC5  0.26631080
## RC4 -0.27470441
## RC3  0.36722816
## RC6 -0.10034163
## RC7 -0.18193663
```



group 0



group 1

```
## [1] 0.9836601
```

The accuracy of LDA in splitting cervical cancer patients is 98%. Unfortunately the class is very imbalanced with 90%+ of the subjects not having cervical cancer, for this type of analysis accuracy is not viable so precision and recall will be used.

```
#prediction train table
t = table(train$Cervical_cancer, predTrain$class)
t
```

```
##
##       0   1
##   0 590   4
##   1   6  12
```

```
# Precision: tp/(tp+fp):
print(paste('Precision', t[1,1]/sum(t[1,1:2])))
```

```
## [1] "Precision 0.993265993265993"
```

```
# Recall: tp/(tp + fn):
print(paste('Recall', t[1,1]/sum(t[1:2,1])))
```

```
## [1] "Recall 0.98993288590604"
```

```
# F-Score: 2 * precision * recall /(precision + recall):
print(paste('FScore', (2 * t[1,1]/sum(t[1,1:2]) * t[1,1]/sum(t[1:2,1])) / (t[1,1]/sum(t[1,1:2]) + t[1,1
```

```
## [1] "FScore 0.991596638655462"
```

```
#testing accuracy
mean(predTest$class==test$Cervical_cancer)
```

```
## [1] 0.974026
```

Confusion Matrix

```r
#prediction test table
z = table(test$Cervical_cancer, predTest$class)
z
```

```
##
##       0   1
##   0 147   1
##   1   3   3
```

LDA with the PCA transformed data set had accuracies of 97.8% and 99.8% on training and accuracy. With a precision of 1, and recall of 0.9801 on the testing data. Precision is defined as number of true positives divided by number of true positives and number of false positives. In this case that's the number of people with cervical cancer divided by all everyone the model said had cervical cancer. Recall on the other hand examines the ability to find all relevant cervical cancer instances. This model does an excellent job of identifying all relevant instances (recall) and returning only the relevant instances (precision). From the histogram binning we can see the model does a good job of splitting the data after "2" on the groupings, but in between 0-2 there is some overlapping between the two groups. The extremely high scores from training and testing may indicate our model is overfitting and in future analysis we can take steps to reduce it.

```r
# Logistic regression
model = glm(Cervical_cancer ~ .,
            family=binomial(link='logit'),
            data=train)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
summary(model)
```

```
##
## Call:
## glm(formula = Cervical_cancer ~ ., family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.23743  -0.06580  -0.02618  -0.00578   2.22246
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -10.5684     2.5927  -4.076 4.58e-05 ***
## RC2          -6.6146     4.2006  -1.575   0.1153
## RC1           1.3613     0.3084   4.414 1.01e-05 ***
## RC5           0.3353     0.5244   0.639   0.5226
## RC4          -7.8587     3.2284  -2.434   0.0149 *
## RC3          -0.6689     0.6370  -1.050   0.2937
## RC6          -0.8623     0.7282  -1.184   0.2364
## RC7          -0.2089     0.8174  -0.256   0.7983
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 162.414  on 611  degrees of freedom
## Residual deviance:  31.362  on 604  degrees of freedom
## AIC: 47.362
##
```

```
## Number of Fisher Scoring iterations: 11
```

```
pred = predict(model, newdata=test, type='response')
```

```
# Let's use the .5 cutoff to classify them
pred = ifelse(pred > 0.5, 1, 0)
table(pred, test$Cervical_cancer)  # Look at false positives and false negatives
```

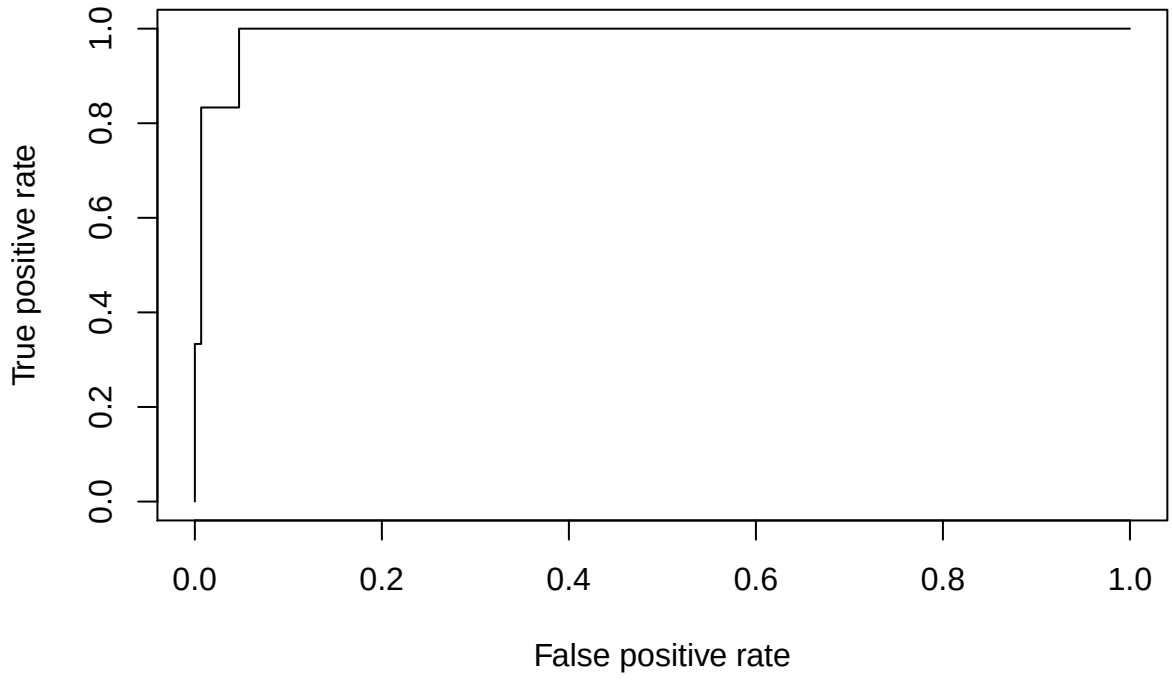
```
##
## pred   0   1
##    0 147   1
##    1   1   5
```

```
misClasificError = mean(pred != test$Cervical_cancer)
print(paste('Accuracy',1-misClasificError))
```

```
## [1] "Accuracy 0.987012987012987"
```

Examining the confusion matrix, we can see the first row is about the non-cervical cancer cases: 147 patients were correctly classified without cervical cancer (called true negatives) and 2 were wrongly classified as having cervical cancer (false negatives). The second row is about the cervical cancer patients: 1 patient was wrongly classified as having cervical cancer (false positives) and 4 were correctly classified as having it (true positives).

```
# Let's look at the ROC curve
p = predict(model, newdata=test, type="response")
pr = prediction(p, test$Cervical_cancer)
prf = performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)
```




```
# Compute the area under the curve
auc = performance(pr, measure = "auc")
auc = auc@y.values[[1]]
auc
```

```
## [1] 0.9887387
```

Using the PCA to predict new dataset and running logistic regression yielded a classification accuracy of 99%. PC1, PC2 and PC4 were significant at a 0.1 level with PC1 and PC4 being significant at 0.05. Yielded an AUC score of 0.9883.