



Multicollinearity: diagnostics and PCA as a method of handling

**Πολυσυγγραμμικότητα: διαγνωστικοί έλεγχοι και η παλινδρόμηση με κύριες
συνιστώσες ως μέθοδος χειρισμού της.**

Ιωάννης Κουτσογιώργος

ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ

Απ. Μπατσίδης: Επίκουρος Καθηγητής Τμ. Μαθηματικών Παν/μιου Ιωαννίνων

**Ι. Στεφανίδης: Καθηγητής Παθολογίας/Νεφρολογίας, Ιατρική Σχολή,
Πανεπιστήμιο Θεσσαλίας**

**Χρ.Δοξάνη: Επιστημονικός Συνεργάτης στη Γενετική Φαρμακοεπιδημιολογία,
Πανεπιστήμιο Θεσσαλίας**

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΛΑΡΙΣΑ 2017

ABSTRACT

Multicollinearity refers to a situation in which two or more explanatory variables in a multiple regression or logistic regression model are highly linearly related. The purpose of this diploma thesis is to find out two things. The first is to verify through some diagnostic tests that our data set is suffered from multicollinearity and the second one to perform principal component analysis to figure out if this phenomenon is eliminated. The results showed that not only multicollinearity of the data set was eliminated but also the predictive model after PCA was better than the one before PCA.

ΠΕΡΙΛΗΨΗ

Η πολυσυγγραμικότητα αναφέρεται σε μία κατάσταση όπου δύο ή περισσότερες επεξηγηματικές μεταβλητές ενός πολλαπλού μοντέλου παλινδρόμησης ή ενός μοντέλου λογιστικής παλινδρόμησης έχουν υψηλή γραμμική συσχέτιση. Ο σκοπός αυτής της διπλωματικής εργασίας είναι να ερευνήσει δύο πράγματα. Το πρώτο είναι να επαληθεύσει μέσω κάποιων διαγνωστικών ελέγχων το γεγονός ότι το δικό μας σύνολο δεδομένων υποφέρει από πολυσυγγραμικότητα και το δεύτερο είναι να εφαρμόσει ανάλυση κύριων συνιστωσών για να ελέγξει αν αυτό το φαινόμενο έχει εξαλειφθεί. Τα αποτελέσματα έδειξαν ότι όχι μόνο εξαλείφθηκε η πολυσυγγραμικότητα στο σύνολο δεδομένων αλλά και το μοντέλο πρόβλεψης μετά την ανάλυση κύριων συνιστωσών ήταν καλύτερο από το μοντέλο πριν την ανάλυση κύριων συνιστωσών.

1. Introduction

Multicollinearity refers to a situation in which two or more explanatory variables in a multiple regression model or logistic regression model are highly linearly related. From this problem our data set (Wisconsin Diagnostic Breast Cancer) was suffered, as some diagnostic tests verified it. The next step was to perform principal component analysis so to eliminate multicollinearity. The results showed that not only multicollinearity of the data set was eliminated but also the predictive model after PCA was better than the one before PCA.

The rest of this diploma thesis is organized as follows. In the second paragraph the meaning of multicollinearity was described as well as the five methods of multicollinearity detection.

In the third paragraph the ways for dealing with multicollinearity were presented, focusing on the principal component analysis. There were described analytical not only the meaning of PCA but also advantages and disadvantages of this technique.

In the last part of diploma thesis a statistical analysis (through R programming) of the chosen data set was performed presenting all outcome and results from that analysis.

2. Multicollinearity

2.1 Definition of Multicollinearity

According to reference [3] multicollinearity refers to a situation in which two or more explanatory variables in a multiple regression model or logistic regression model are highly linearly related. We have perfect multicollinearity if the correlation between two independent variables is equal to 1 or -1 . In practice, we rarely face perfect multicollinearity in a data set. More commonly, the issue of multicollinearity arises when there is an approximate linear relationship among two or more independent variables.

2.2 Methods for multicollinearity detection

According to references [1], [2], [5], there are several methods to detect multicollinearity, which are:

- Variance Inflation Factor
- Tolerance
- Condition Number
- Condition Index
- Variance decomposition-proportion

2.2.1 Variance Inflation Factor (VIF)

Variance inflation factors measure the inflation in the variances of the parameter estimates due to collinearities that exist among the predictors. It is a measure of how much the variance of the estimated regression coefficient β_k is “inflated” by the existence of correlation among the predictor variables in the model. A VIF of 1 means that there is no correlation among the k th predictor and the remaining predictor variables, and hence the variance of β_k is not inflated at all. The general rule of thumb is that VIFs exceeding 4 warrant further investigation, while VIFs exceeding 10 are signs of serious multicollinearity requiring correction.

2.2.2 Tolerance

Tolerance is a measure of collinearity. The variable’s tolerance is $1-R_j^2$ (where R_j^2 is the coefficient of determination of a regression of explanator j on all the other explanators) . A small tolerance value indicates that the variable under consideration is almost a perfect linear combination of the independent variables already in the equation and that it should not be added to the regression equation. All variables involved in the linear relationship will have a small tolerance. Some suggest that a

tolerance value less than 0.2 or 0.1 should be investigated further. If a low tolerance value is accompanied by large standard errors and non-significance, multicollinearity may be an issue.

2.2.3 Condition Number

Another measure of the overall multicollinearity of the variables can be obtained by computing the condition number (CN) of the correlation matrix, defined by the ratio of the largest Eigenvalue to smallest Eigenvalue (Chatterjee and Hadi, 2006). The condition number will always be greater than 1. A large condition number (larger than 15) indicates evidence of collinearity.

2.2.4 Condition Index

Most multivariate statistical approaches involve decomposing a correlation matrix into linear combinations of variables. The linear combinations are chosen so that the first combination has the largest possible variance (subject to some restrictions we won't discuss), the second combination has the next largest variance, subject to being uncorrelated with the first, the third has the largest possible variance, subject to being uncorrelated with the first and second, and so forth. The variance of each of these linear combinations is called an eigenvalue. Collinearity is spotted by finding 2 or more variables that have large proportions of variance (.50 or more) that correspond to large condition indices. A rule of thumb is to label as large those condition indices in the range of 30 or larger.

2.2.5 Variance decomposition-proportion

The variance-decomposition proportions (VD) are the variance proportions of the i -th variable attributable to the j -th eigenvalue. No variable should attribute more than 0.5 to any one eigenvalue (Dormann et al., 2012).

3. Principal Component Analysis (PCA) as a method of handling with multicollinearity

Depending on what the source of multicollinearity is, the solutions will vary. If the multicollinearity has been created by the data collection, collect additional data over a wider X-subspace. If the choice of the linear model has increased the multicollinearity, simplify the model by using variable selection techniques. If an observation or two has induced the multicollinearity, remove those observations. Above all, use care in selecting the variables at the outset. When these steps are not possible, you might try:

- ridge regression and
- principal component analysis (PCA).

In this diploma thesis our interest will be focused on PCA. PCA meaning, advantages and disadvantages are according to reference [4].

3.1 Principal Component Analysis (PCA)

PCA is a commonly used data reduction technique (Abdi and Williams 2010). This method seeks to find linear combinations of the predictors, known as principal components (PCs), which capture the most possible variance. The first PC is defined as the linear combination of the predictors that captures the most variability of all possible linear combinations. Then, subsequent PCs are derived such that these linear combinations capture the most remaining variability while also being uncorrelated with all previous PCs. Mathematically, the j th PC can be written as:

$$PC_j = (a_{j1} \times \text{Predictor 1}) + (a_{j2} \times \text{Predictor 2}) + \dots + (a_{jP} \times \text{Predictor P}).$$

P is the number of predictors. The coefficients $a_{j1}, a_{j2}, \dots, a_{jP}$ are called component weights and help us understand which predictors are most important to each PC.

3.2 Advantages

The primary advantage of PCA, and the reason that it has retained its popularity as a data reduction method, is that it creates components that are uncorrelated. Some predictive models prefer predictors to be uncorrelated (or at least low correlation) in order to find solutions and to improve the model's numerical stability. PCA preprocessing creates new predictors with desirable characteristics for these kinds of models.

3.3 Disadvantages

PCA seeks predictor-set variation without regard to any further understanding of the predictors (i.e., measurement scales or distributions) or to knowledge of the modeling objectives (i.e., response variable). Hence, PCA can generate components that summarize characteristics of the data that are irrelevant to the underlying structure of the data and also to the ultimate modeling objective.

Because PCA seeks linear combinations of predictors that maximize variability, it will naturally first be drawn to summarizing predictors that have more variation. If the original predictors are on measurement scales that differ in orders of magnitude, then the first few components will focus on summarizing the higher magnitude predictors, while latter components will summarize lower variance predictors. This means that the PC weights will be larger for the higher variability predictors on the first few components. In addition, it means that PCA will be focusing its efforts on identifying the data structure based on measurement scales rather than based on the important relationships within the data for the current problem.

The second caveat of PCA is that it does not consider the modeling objective or response variable when summarizing variability. Because PCA is blind to the response, it is an *unsupervised technique*. If the predictive relationship between the predictors and response is not connected to the predictors' variability, then the derived PCs will not provide a suitable relationship with the response.

4. Practical application and results

The data set I chose is called Wisconsin Diagnostic Breast Cancer (WDBC) and it was downloaded from reference [6]. Except from attribute information which are ID number and Diagnosis (dependent variable) (M = malignant, B = benign) there are also ten real-valued features (independent variables) which are computed for each cell nucleus:

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

These features were 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.

Also, 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.

This data set consists of 569 observations of 32 variables as shown below:

This study of the data set consists of these stages:

- Test for multicollinearity using diagnostic tests
- If multicollinearity appears, use PCA to eliminate phenomenon
- Test if the model, after PCA handling, has better predictive power than the model before it

4.1 Test for multicollinearity using diagnostic tests

In this paragraph we will use the diagnostics in order to check if multicollinearity is present.

1. Variance Inflation Factor (VIF)

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source
Console C:\Users\basilio\Documents\...

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> cancer<-read.csv("data.csv")
>
> cancer$K~MALL
> dim(cancer)
[1] 569 32
> cancer$id<-NULL
> model <- glm(diagnosis ~.,family=binomial(link='logit'),data=cancer)
warning messages:
1: glm.fit: algorithm did not converge
2: glm.fit: Fitted probabilities numerically 0 or 1 occurred
> library(car)
> vif(model)
radius_mean          Texture_mean          perimeter_mean
4318063.739          140816.298          1801913.723
area_mean            smoothness_mean          compactness_mean
6331653.418          213016.609          413839.479
concavity_mean       concave.points_mean       symmetry_mean
185593.174           192381.200          11851.056
fractal_dimension_mean radius_se          texture_se
3513.136             610335.622          350773.141
perimeter_se         area_se           smoothness_se
49276.836            1109444.172          41333.693
compactness_se       concavity_se       concave.points_se
673636.453           778242.166          1574935.418
symmetry_se          fractal_dimension_se radius_worst
24678.711            463075.232          3511808.059
texture_worst        perimeter_worst     area_worst
823081.426           617752.395          4767645.776
smoothness_worst     compactness_worst   concavity_worst
56738.790            91238.174           1705825.343
concave.points_worst symmetry_worst      fractal_dimension_worst
561067.026           10400.330           103961.536

```

Variance inflation factors measure the inflation in the variances of the parameter estimates due to collinearities that exist among the predictors. As we see VIFs are exceeding the number 10 so it's a sign of serious multicollinearity.

2. Condition Number


```

> cancer<-read.csv("data.csv")
> cancer[1:5,]
  radius_mean  texture_mean  perimeter_mean
1 4518663.730  140850.298  1091015.723
2 6314531.030  213020.000  413630.479
3 105593.174  182381.200  11851.056
4 3513.136  610325.622  356773.141
5 49276.036  1108444.172  41533.493
6 473836.433  778242.166  1374953.418
7 24874.711  463075.212  331808.058
8 823681.626  617732.395  870685.776
9 16734.790  92236.174  1703825.143
10 561962.026  10409.339  191901.538
> fit=glm(x=cancer[,1:10], y=cancer[,11], family=binomial(link=logit), data=cancer)
Writing messages to stdout:
1: glm.fit: algorithm did not converge
2: glm.fit: fitted probabilities numerically 0 or 1 occurred
> fit$coefficients
(1) 61907723
  
```

Condition's number value is 61907723 which is above 15. This difference in values indicates multicollinearity problem.

3. Condition Index

```

> library(base)
> kappa2(model)
[1] 61907723
> library(perturb)
> test=colSubst(model)
> test$condindx
cond_index
1 1.000000
2 4.343154
3 5.154472
4 6.659817
5 12.323639
6 13.782767
7 14.320395
8 15.528468
9 16.738827
10 19.380883
11 20.835157
12 27.753718
13 29.537770
14 37.647941
15 41.242122
16 46.321176
17 51.659064
18 56.393252
19 62.934581
20 68.347687
21 74.070762
22 82.573455
23 94.591543
24 105.988837
25 127.944872
26 142.593022
27 151.738086
28 284.650849
29 404.182319
30 596.903329
31 1805.649970
  
```

As we see there are values between 30-100 (14 through 23) which indicate moderate multicollinearity problem and there are values above 100 (24 through 31) which indicate serious multicollinearity problem.

4. Variance decomposition-proportion

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
V. Admin
Source on Save
Console
C:\Users\...
C# 404.182319
30 596.961529
31 1805.649070
= Test101
Intercept radius_mean texture_mean perimeter_mean area_mean smoothness_mean compactness_mean concavity_mean concave_points_mean symmetry_mean fractal_dimension_mean
[1] 6.773165e-07 2.026900e-08 5.039156e-08 2.278301e-08 7.858513e-07 3.288301e-06 5.181019e-06 6.874284e-06 7.329379e-06 5.453966e-06 1.008761e-06
[2] 1.523472e-03 3.932052e-09 6.370618e-05 3.720148e-10 1.382235e-05 4.758083e-05 1.045871e-05 3.880578e-04 3.801403e-04 1.034226e-04 2.376857e-03
[3] 5.247084e-06 3.708880e-07 4.704155e-05 3.702015e-07 2.579327e-05 1.087121e-05 7.766950e-05 2.888275e-04 2.817070e-06 1.582897e-05 6.828674e-07
[4] 6.853471e-07 1.082064e-07 6.817118e-08 1.794912e-07 2.835238e-06 7.736873e-06 9.337901e-05 8.106301e-05 1.475954e-04 7.521257e-06 5.021131e-07
[5] 1.758801e-05 2.624384e-06 7.247081e-04 2.496951e-06 1.930158e-04 1.343824e-04 2.035866e-03 1.749448e-05 4.050545e-04 1.573574e-04 1.197453e-05
[6] 5.011717e-05 1.720883e-06 1.783910e-03 1.688985e-06 6.359350e-05 2.827995e-04 1.338575e-05 1.038021e-03 2.111051e-05 2.435713e-04 7.341371e-05
[7] 1.421381e-05 7.540807e-09 5.615256e-04 1.124270e-10 1.798463e-06 1.817364e-04 2.060652e-04 3.471788e-03 6.693680e-03 4.189071e-05 1.780796e-05
[8] 1.976882e-03 1.942888e-08 1.934420e-04 7.298795e-08 3.611338e-05 1.400138e-06 4.090628e-04 7.201888e-05 2.342616e-03 1.127301e-03 3.843228e-06
[9] 1.207914e-03 1.830800e-08 1.104121e-04 2.411109e-06 2.247667e-04 4.331372e-04 5.454732e-04 4.224237e-04 1.641607e-03 1.045350e-03 5.480099e-05
[10] 8.473035e-08 2.373229e-07 3.552127e-04 3.390041e-07 1.722182e-04 1.320839e-06 3.303312e-05 9.444416e-05 1.938523e-03 1.639300e-06 2.787991e-06
[11] 2.342592e-05 9.571078e-07 4.938631e-04 1.110883e-06 3.963385e-04 1.261883e-03 3.843594e-04 6.811718e-04 2.284440e-03 1.786953e-01 3.177044e-04
[12] 6.115133e-04 1.047644e-03 3.906645e-03 1.343705e-05 1.450199e-04 3.381034e-03 1.137460e-02 2.835048e-03 1.355026e-04 2.118774e-03 8.843319e-04
[13] 3.865888e-07 1.236914e-07 4.306892e-02 5.441728e-10 6.759761e-05 3.699374e-05 5.850412e-03 1.780390e-02 5.999519e-03 1.154732e-04 1.700108e-05
[14] 1.419868e-04 8.830238e-06 6.208970e-05 1.104258e-05 3.837337e-05 1.740713e-03 1.390950e-03 1.564788e-03 2.854916e-03 7.274518e-03 2.493796e-05
[15] 8.066592e-04 2.783111e-05 1.668357e-03 3.322374e-05 1.870083e-03 2.924734e-04 9.031914e-03 1.545304e-03 3.460454e-04 1.827905e-02 4.085408e-04
[16] 4.022928e-04 5.170352e-06 8.744488e-08 3.227589e-06 6.870660e-06 2.191458e-03 7.298732e-02 7.041898e-03 1.361171e-02 8.701208e-02 5.479324e-04
[17] 3.767462e-04 6.522773e-07 3.947867e-04 2.462894e-06 1.657691e-03 2.533903e-02 2.347665e-02 8.437236e-03 3.488196e-02 8.801771e-02 1.662352e-04
[18] 1.729344e-04 4.089618e-05 1.713063e-03 4.343642e-05 3.029809e-03 4.635102e-03 1.066998e-03 2.768087e-02 1.465393e-02 8.708257e-03 1.590918e-01
[19] 2.816716e-04 3.324383e-03 1.085713e-02 2.242214e-05 5.712368e-04 1.468550e-02 4.518248e-04 2.042802e-02 4.953599e-03 8.477940e-03 4.664014e-03
[20] 4.146772e-03 1.099187e-05 5.135289e-04 2.444008e-05 3.225465e-03 1.818865e-02 8.277170e-02 4.464728e-02 6.442420e-02 1.450589e-02 1.080007e-02
[21] 3.310678e-03 8.415098e-04 7.181125e-03 8.057137e-06 3.542758e-03 4.585948e-02 8.724800e-02 1.120591e-02 3.620227e-02 1.265276e-02 7.047357e-03
[22] 1.358387e-03 5.206286e-05 4.412784e-02 6.218372e-05 2.087761e-03 1.476688e-02 3.219072e-03 3.904089e-01 2.533225e-01 1.574309e-01 1.358552e-03
[23] 2.867140e-03 2.292216e-06 1.856085e-02 2.865178e-06 5.288816e-03 2.475569e-02 9.637115e-02 1.850582e-01 9.803319e-02 2.110741e-01 1.542395e-03
[24] 1.538428e-02 5.320894e-05 7.860134e-03 6.140953e-05 6.188114e-02 7.619747e-03 2.403700e-03 1.084799e-01 2.338005e-01 6.272699e-02 2.125443e-04
[25] 1.702049e-05 2.494680e-05 5.419647e-01 3.155646e-05 1.755330e-03 1.335622e-01 3.252021e-03 2.278012e-02 6.232044e-02 1.580755e-01 5.696146e-03
[26] 4.020422e-02 1.640027e-04 3.321401e-02 2.276278e-04 3.877707e-02 9.333413e-02 1.238070e-03 3.354412e-02 3.142821e-02 8.382863e-01 1.945681e-01
[27] 7.980705e-01 2.886440e-04 2.395305e-03 2.392529e-04 1.348215e-02 9.763893e-03 2.275189e-03 3.864830e-04 1.144503e-05 1.009689e-02 7.340878e-01
[28] 3.400703e-02 1.067495e-02 9.602561e-04 6.554039e-03 2.493848e-01 3.222782e-04 8.005331e-03 1.384234e-02 1.384523e-03 2.625161e-05 4.108607e-03
[29] 8.465672e-02 1.840788e-02 1.272639e-02 4.473050e-02 5.792310e-01 3.098808e-03 5.154311e-02 4.167889e-03 3.986038e-04 1.613839e-02 1.538374e-02
[30] 9.101913e-03 9.736573e-01 5.861735e-05 9.478895e-01 2.171109e-02 2.173449e-02 2.964350e-01 6.716895e-02 1.170190e-04 2.667463e-01 1.028865e-02
radius_se texture_se perimeter_se area_se smoothness_se compactness_se concavity_se concave_points_se symmetry_se fractal_dimension_se radius_worst texture_worst
[1] 4.839121e-06 4.432415e-05 3.388969e-06 1.187393e-05 4.283420e-05 2.394378e-05 2.799438e-05 2.291406e-05 3.112223e-05 3.467438e-05 1.335760e-07 3.669378e-06
[2] 1.384898e-04 1.316559e-03 1.702150e-04 1.496404e-03 1.397966e-03 1.687406e-06 1.050451e-04 1.235532e-06 7.807763e-04 2.796771e-04 2.436287e-08 4.720149e-05
[3] 8.064001e-05 8.539957e-05 6.369932e-05 6.119387e-04 1.789819e-05 2.888474e-03 6.344671e-03 5.313237e-04 3.763198e-06 4.316771e-03 2.706112e-06 3.500542e-05
[4] 7.036558e-04 4.599101e-03 7.581265e-04 2.789260e-03 3.837466e-03 4.870120e-04 1.493703e-03 6.483106e-04 1.907758e-03 1.174239e-03 1.183320e-06 3.044083e-05
[5] 4.964729e-04 4.316915e-03 8.303981e-04 1.954458e-03 6.517688e-03 6.402117e-04 4.457305e-02 4.437239e-03 7.924104e-03 2.753233e-05 1.235296e-05 4.705482e-04

```

There are at least two regression coefficients with variance-decomposition proportion bigger than 50%. This fact leads us to multicollinearity problem.

5. Tolerance

This diagnostic test is not applicable to specific data set as the dependent variable (diagnosis) is binary.

4.2 PCA as a method of handling multicollinearity

	PC1	PC2	PC3	PC4	PC5
texture_mean	0.1189244	0.2381732	0.008531249	0.011488962	0.017786514
texture_worst	-0.11072458	0.25970088	0.064348903	-0.023030661	0.048488910
perimeter_mean	0.22751729	0.215183161	-0.084813210	0.041981989	-0.017316661
area_mean	-0.27089499	0.23107611	0.028889326	0.053433785	-0.030331231
compactness_mean	-0.18138881	-0.186115021	-0.10249384	0.09982745	0.165886128
compactness_worst	-0.23828531	-0.151801618	-0.074891371	0.051794381	-0.031705937
concavity_mean	-0.25882088	-0.060165484	0.007738838	0.019122713	-0.088873412
concavity_worst	-0.26481376	0.074767108	-0.025583541	0.095335944	0.041881031
symmetry_mean	-0.11828986	0.18026772	-0.08023886	0.062149884	0.105381428
fractal_dimension_mean	-0.08438131	-0.189573471	-0.023374930	0.049586585	0.044428360
radius_se	-0.20587878	0.105521312	0.248881307	0.041981212	0.134856496
texture_se	-0.02742953	0.089978682	0.274835903	-0.230013128	0.185839338
perimeter_se	-0.23127592	0.084657234	0.266445307	0.048900415	0.120890220
area_se	-0.20288884	0.152292628	0.218888126	0.128507199	0.127574452
compactness_se	-0.01451245	0.29430831	0.168838179	0.04664138	0.232856716
concavity_se	-0.1939485	0.23271888	0.134778718	-0.027448917	-0.278888136
concave_perim_se	-0.13158979	0.187207283	-0.174481743	0.001316889	0.103482091
concave_perim_worst	-0.18317161	0.18121184	-0.224817407	0.273061111	0.195588889
symmetry_se	-0.04248847	-0.18848850	0.288888392	0.044073351	0.252868703
fractal_dimension_se	-0.11058837	0.28009107	0.213393764	0.15398479	0.161297438
radius_worst	-0.22788093	0.218861378	-0.047588900	0.01417248	0.00488302
texture_worst	-0.10448813	0.08546284	-0.082287673	-0.02895881	0.08388101
perimeter_worst	-0.11883888	0.09878478	-0.048548508	0.113800784	-0.007454131
area_worst	-0.12487832	0.21911818	-0.011862318	0.075887188	0.027388921
compactness_worst	-0.12785296	0.172984312	-0.158787815	0.017652128	0.124834441
concavity_worst	-0.13088588	-0.182092174	-0.216675825	-0.091328813	-0.171888107
concave_perim_worst	-0.01876713	-0.097964104	-0.173857733	-0.07051148	-0.188516225
concave_perim_mean	-0.15888887	0.08825725	-0.123488916	0.006688888	-0.04110086
symmetry_worst	-0.12298886	-0.141885349	-0.273128847	-0.036738881	0.244538863
fractal_dimension_worst	-0.12188886	-0.273398848	-0.210781813	-0.07702478	-0.084421021

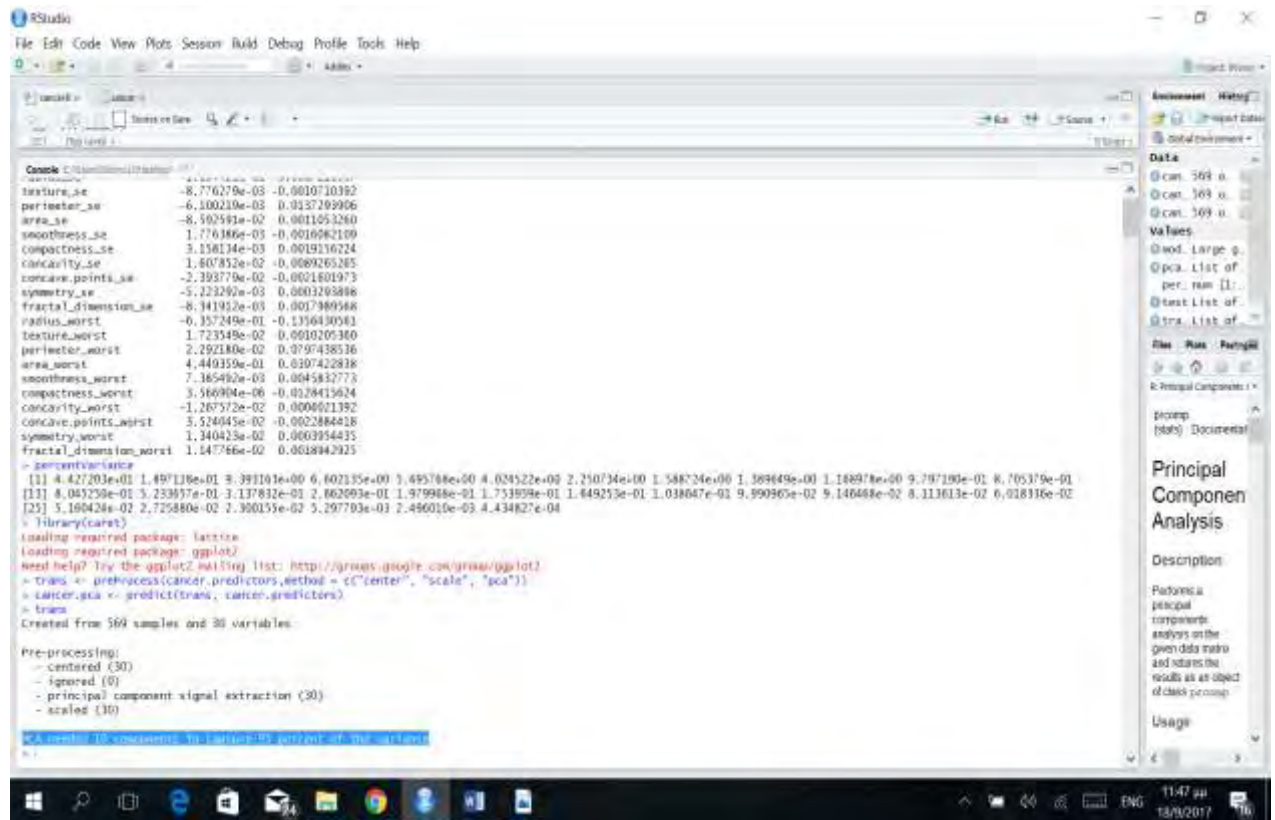
There are 30 PCs that are shown above. PC1 is defined as the linear combination of the predictors that captures the most variability of all possible linear combinations. Then, subsequent PCs are derived such that these linear combinations capture the most remaining variability while also being uncorrelated with all previous PCs.

The percentage of variability that is captured by each PC is:

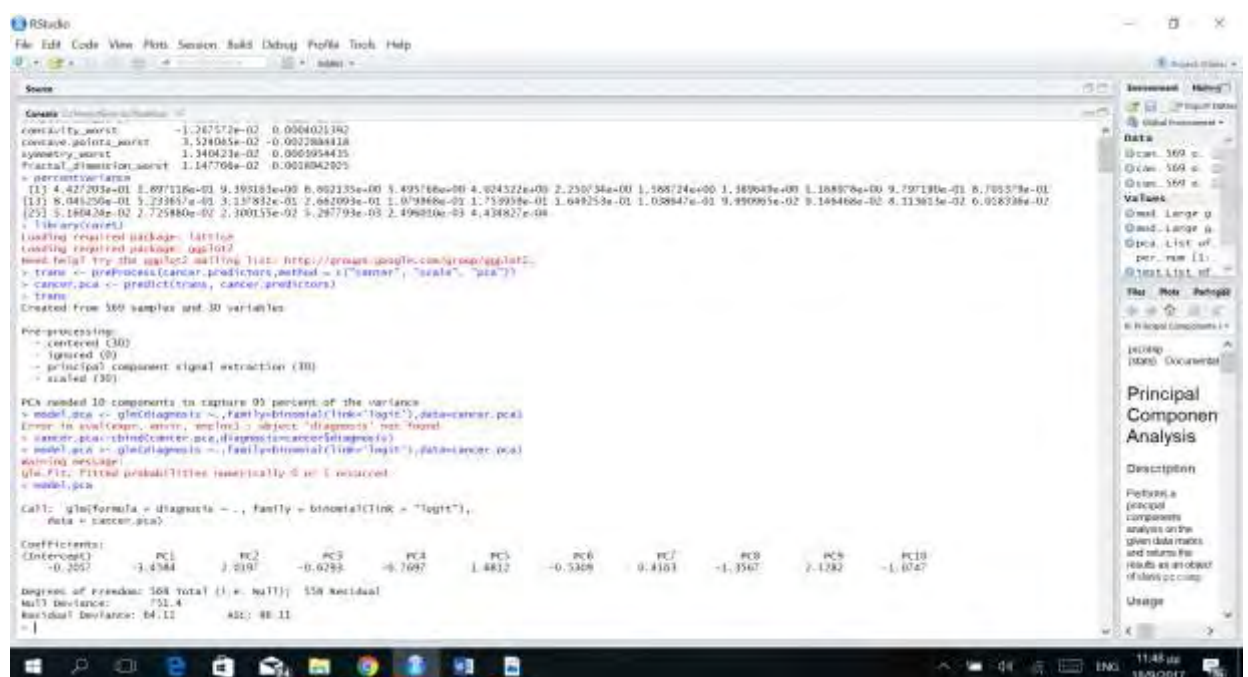
	PC1	PC2	PC3	PC4	PC5
area_worst	-0.03828935	0.231358525	0.237162856	2.389683e-01	
compactness_worst	-0.04786678	0.012692464	-0.048853588	-1.155248e-01	
concavity_worst	-0.02438884	-0.100463424	-0.079505414	-0.869812e-02	
concave_perim_worst	0.11377034	0.246833781	-0.142805803	1.768978e-02	
symmetry_worst	0.28110034	-0.15574307	0.216801388	2.247387e-02	
fractal_dimension_worst	0.04529982	0.028184296	0.022700444	4.020481e-03	
radius_mean	2.114605e-01	0.7824149920			
perimeter_mean	3.838261e-01	-0.6888695883			
area_mean	-4.727949e-01	-0.0329473882			
smoothness_mean	-3.434667e-03	-0.0048474577			
compactness_mean	-4.101677e-02	0.0486741883			
concavity_mean	-1.052479e-02	0.0231286681			
concave_perim_mean	-4.266494e-03	-0.0015772633			
symmetry_mean	-7.588862e-03	-0.0012803794			
fractal_dimension_mean	7.301433e-03	-0.0047568848			
radius_se	1.184421e-01	-0.0087110937			
perimeter_se	-6.776279e-03	-0.0016710392			
area_se	-6.180021e-02	0.0137290366			
smoothness_se	-8.292191e-02	0.0011053260			
compactness_se	1.778389e-03	-0.0018082160			
concavity_se	3.158134e-03	0.0018156224			
concave_perim_se	1.667852e-02	-0.0089265285			
concave_perim_worst	-2.39379e-02	-0.0021601373			
symmetry_se	-5.222928e-02	0.003791888			
fractal_dimension_se	-8.345012e-03	0.0017889568			
radius_worst	-6.357249e-01	-0.135430581			
perimeter_worst	1.723149e-02	0.0010203380			
area_worst	2.292180e-02	0.0797489536			
smoothness_worst	4.848959e-01	0.087423818			
compactness_worst	3.385492e-03	0.004562773			
concavity_worst	-3.588804e-06	-0.0128415624			
concave_perim_worst	-1.267572e-02	0.0094021182			
symmetry_worst	3.524045e-02	-0.0022884418			
fractal_dimension_worst	1.349423e-02	0.0031854415			

4.3 Test if the model, after PCA handling, has better predictive power than the model before it.

PCA needed 10 components to capture 95% of the variance.



Model after PCA



Model before PCA

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help

Console
data = cancer.pca

Coefficients:
(Intercept)      PC1      PC2      PC3      PC4      PC5      PC6      PC7      PC8      PC9      PC10
-0.2037      -3.4384      2.0197      -0.0293      -0.7697      1.6812      -0.3309      0.4163      -1.3347      2.1282      -1.0747

Degrees of Freedom: 568 Total (i.e. Null); 558 Residual
Null Deviance: 751.4
Residual Deviance: 84.11      AIC: 86.11
= cancer=read.csv("data.csv")
= cancer$y=MLL
= cancer$y=NULL
= model <- glm(cancer.y ~ ., family=binomial(link='logit'), data=cancer)
Warning messages:
1: glm.fit: algorithm did not converge
2: glm.fit: fitted probabilities numerically 0 or 1 occurred
= model

Call: glm(formula = diagnosis ~ ., family = binomial(link = "logit"),
data = cancer)

Coefficients:
(Intercept)      radius_mean      texture_mean      perimeter_mean
-2.661e+06      2.427e+08      1.958e+05      1.473e+08
area_mean      smoothness_mean      compactness_mean      concavity_mean
-1.301e+05      -1.525e+04      -6.428e+06      1.042e+08
concave.points_mean      symmetry_mean      fractal_dimension_mean      radius_se
-1.216e+07      4.045e+07      -4.233e+07      3.328e+07
texture_se      perimeter_se      area_se      smoothness_worst
4.368e+06      1.752e+06      -8.393e+05      7.482e+08
compactness_worst      concavity_worst      concave.points_worst      symmetry_worst
-1.773e+08      1.329e+04      -1.200e+09      2.890e+08
fractal_dimension_worst      radius_worst      texture_worst      perimeter_worst
1.512e+09      -6.130e+09      -5.632e+05      -3.538e+05
area_worst      smoothness_worst      compactness_worst      concavity_worst
8.850e+04      -2.152e+07      4.909e+06      -3.028e+07
concave.points_worst      symmetry_worst      fractal_dimension_worst
1.431e+08      -2.474e+07      -1.698e+07

Degrees of Freedom: 568 Total (i.e. Null); 558 Residual
Null Deviance: 751.4
Residual Deviance: 32010      AIC: 12070
= |
  
```

In conclusion, we ended up with 2 results:

- I. We eliminate multicollinearity of the data set, through principal component analysis (PCA)
- II. The predictive model after PCA is better than the one before PCA. AIC metric indicates that. In the model after PCA AIC is 86.11 and in the other is 32070. And we know that smaller AIC number leads to better model.

5. R code

```
cancer<-read.csv("data.csv")
names(cancer)
cancer$X<-NULL
cancer$id<-NULL
model <- glm(diagnosis ~.,family=binomial(link='logit'),data=cancer)
#multicollinearity check with vif
library(car)
vif(model)
#multicollinearity check with Condition indexes and variance decomposition
proportions
library(perturb)
test<-colldiag(model)
#condition indexes
test$condindx
#variance decomposition proportions
test$pi
#multicollinearity check with condition number
library(base)
kappa(model)
cancer.predictors<-cancer[,-1]
pca.cancer <- prcomp(cancer.predictors,center = TRUE, scale. = TRUE)
percentVariance <- pca.cancer$sd^2/sum(pca.cancer$sd^2)*100
percentVariance
library(caret)
trans <- preProcess(cancer.predictors,method = c("center", "scale", "pca"))
cancer.pca <- predict(trans, cancer.predictors)
cancer.pca<-cbind(cancer.pca,diagnosis=cancer$diagnosis)
model.pca <- glm(diagnosis ~.,family=binomial(link='logit'),data=cancer.pca)
```

6. References

1. *Chatterjee and Hadi, 2006*
2. *Dormann et al. (2012)*
3. *Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, "An Introduction to Statistical Learning" 2013.*
4. *Max Kuhn, Kjell Johnson, "Applied Predictive Modeling" 2013.*
5. *O'Brien(2007), Belsley (1991), Farrar, Glauber (1967), Wichers (1975), Kuman (1975) and O'Hagan,McCabe (1975), "Articles of Wikipedia"*
6. *Website "www.kaggle.com"*