

#### ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ ΤΜΗΜΑ ΙΑΤΡΙΚΗΣ



ΠΜΣ «Μεθοδολογία Βιοϊατρικής Έρευνας, Βιοστατιστική και Κλινική Βιοπληροφορική»

#### Multicollinearity: diagnostics and PCA as a method of handling

Πολυσυγγραμικότητα: διαγνωστικοί έλεγχοι και η παλινδόμηση με κύριες συνιστώσες ώς 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 πολυσυγραμικότητα 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 άν αυτό το φαινόμενο έχει εξαλειφθεί. Τα αποτελέσματα έδειξαν ÓΤΙ Óχι μόνο εξαλείφθηκε n πολυσυγγραμικότητα στο σύνολο δεδομένων άλλα και το μοντέλο πρόβλεψης μετά την ανάλυση κύριων συνιστωσών ήταν καλύτερο από το μοντέλο πρίν 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 <u>Definition of Multicollinearity</u>

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 <u>Methods for multicollinearity detection</u>

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  $\beta k$  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 kth predictor and the remaining predictor variables, and hence the variance of  $\beta k$  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 <u>Tolerance</u>

Tolerance is a measure of collinearity. The variable's tolerance is  $1-Rj^2$  (where  $Rj^2$  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 <u>Condition Number</u>

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 <u>Condition Index</u>

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 <u>Variance decomposition-proportion</u>

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 jth PC can be written as:

 $PCj = (aj1 X Predictor 1) + (aj2 X Predictor 2) + \cdots + (ajP X Predictor P).$ 

P is the number of predictors. The coefficients aj1, aj2,..., ajP are called component weights and help us understand which predictors are most important to each PC.

#### 3.2 <u>Advantages</u>

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 (perimeter^2 / 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:

31		-	Turnel														
2		Later -	() segment													4.	
	at -	diagones.	callion must in	two, meal-	primiter much	arcoad	unitary and	comparison must	ton why with		Symmetry Amale	Distal diversion hands	runtes se	THURSDAY.	perimeter.58	4442.56	-
6	\$12,007		17.998	10.10	102.00	1001.0	10.71842	0.27765	0.10010	8 /4/1/6	1,7478	0.07877	1.0988	4.8053	# 1.001	182.400	1
÷.	\$41512	4	31.578	18.27.	11/100	1126.0	0.00474	0.17264	2 SAMEAG	E8/07/N	6102	0.05887	1 3435	4.5388	1.2888	74.102	i
ŧ.	Arabbeck		18.690	11:25	(30.00	1262.0	1.0005	0.15980	A PATHOR	3.127606	1.2018	0.01806	1.560	1.7864	4.5834	HIN	1
	ALACEAN	M .	11.00	.01.44	77.38	1.664		1.25.00	* 20140	a insidesi	1.2547	0.06766	a seis	17568	2 4210	21.210	
6	Asymptot		22.29	12.10	121.10	1397.0	#10934	8,3,0,00	1/100109	0.001408	8 7304	0.01844	6.7572	4,903	6.4 (10)	11.420	
é.	MITTE		12-458	13.29	82.47	487.1	8.12790	E.17080	\$75.969	0.100920	8.2087	0.07918	1.1.145	8.8902	1.079	27.798	
1	844339	ie.	19.239	19.98	119.00	1040.0	0.09403	- 81.102980	E/12708	2.57422	2.1794	8.03742	1.4487	8.1732	3.1809	- 11.911	
6	84418207	*	11.70	20.61	90.25	177.9	.611800	0.18410.	n sousses	e.distet/	6,2198	0.07451	0.5539	7.3778	1,6365	TH.MO	
		47	78.009	21.82	47.22	192	s.127m	0.18340	# /#180#	0.007520	8.2250	0.07388	8.3063	1 ROUS	2 4008	34.326	
	AC\$27307	44	12.48	28.89	14.47	475.0		0.20080	0.32778.0	a setting	8 2048	0.08243	8.2078	1.3989	21000	23.840	- 4
ñ	8739.26	41	18.0.01	28.20	142.70	147.6	8.04208	Usase	1213991	2.833216	81848	0.01840	3 3795	1.2474	3 2 8 8 2	10.010	
i.	AUR10007		15.789	17.89	117 87	.783.0	8.09711	0.12610	3 200542	a nonze	87847	0.04047	0.5748	1160	7 1847	34.700	1.0
ñ	\$16,228	M	19.779	24.87	107.40	1122.0	1.79745	1.24500	6.208509	3.11160	8.2391	0.07908	1.8515	3.3680	11 8708	118.200	
6	\$46.607	M	10,000	11.85	101.79	1927	1.04400	8.19040	1.00000	1.05/840	1 /M	aataa	1.4617	1,0760	2.9036	RIN	1
ŧ.	Autom	Je .	14.738	22.81	42.60	578.2	ALLER	8.27941	1212000	8.880238	8.2089	0.07682	12107	1 7840	21815	16.211	
ł.	Artmony	44	16.568	37.14	W78	10.2	ATTANC.	0.15940	A163809	Automa	8,2364	4.4/4/7	8.4708	1.8348	2.4794	46.310	
£	*10.016	.64	11.449	20.14	84.71	1942.0	3.08847	8.07/00	3.873938	0.0523396	81186	0.018/1	ealor	7.668	1.1016	11.88	
ł.	84952007		18,779	11.60	108.10	758.8	A.11208	1 25220	81/2208	E ASSESS	8,2184	0.07.04	3 9992	1 8740	1000	34.788	
۰.	\$-3014	4	10.810	77.19	100.00	1260.0	1.00471	0.10270	8.747908	2.094345	¥ /342	0.01301	1.7847	1,0170	1.013	112,400	
1	RIPIZ		/3.548	14.38	17.48	165.3	8.09779	0.08129	a stephet	2,94/912	8.7985	0.01788	1.7659	1.788	21589	23.998	
ŧ.	#110653	4	10.00	1177	41.63	100.0	#16736	6.0780	8.045588	3 831708	82901	0.0681	3/152	8.1477	1.3628	14.879	
	structo	4	8.501	12.64	80.34	273.9	8.16242	U Denau	2.039588	3 220766	1.1813	A pelakts	0.1773	6.8798	1.856	12,000	- 6
10	airrin	41	15.44	11.24	132.00	192.2	.it 19736	e Jezan	1.00708	3.55764	8,2521	3.0940	8-1168	2.0168	2.5822	21.814	
	411329	N.	21.788	/1.11	107.28	1484.0	8.29429	0.16220	10000	1.044.025	81799	-0.01278	0.6917	1.1278	+ 100	11.000	
1	#12552	M .	18,010	21.28	110.00	104.0	0.11218	0.14370	875250	8.891708	E.1991	0.062.01	1.000	2.0077	2.4538	1472 806	
s.	#53527		17.140	18.48	-116.80	9127	A71840	0.22760	1222284	it Avenue	18.8040	0.0411	1,848	0.8700	7.2768	212-400	1
1																	3
	a i a ti d	and taken i						_							_	_	-

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)

to Edit Lodo Yww Plots	Session Build Debug Profile	Tools Help	
		p.	E from these
Source			at
Canada E.Alama Kaneta Daimar			-5
<pre># is a collaborative pro Type 'contributors()' Fo 'citation()' on how to c</pre>	ject with many contributo in sorw information and tite & or & packages in pa	s. Hrating.	
type 'demo()' for some d 'help.start()' for an WT Type 'd()' to guit R.	innos, 'help()' for on-lin ML broader interface to h	rhelp, or lp.	
s cantars-read.cav("data	("van,		
· cancer\$x-MULL			
> dmutametri 11 J69 32 > concertide NutL > concertide NutL > madel ~ e place lagnes is warring messages 10 justit: adporting at 2 glas fit: adporting at 2 hibeary(car) > viffmodel Alloca, J39 area.mean 031063, J39	<pre>ifamily-binumial(link- d not converge b)llties numerically 0 or texture_mean 140816.208 assorbines_mean 213016.609</pre>	legit").data-cariar) 1.occurred perimeter_mean 1801913.733 campatines_mean 431589.479	
concavity_mean 105593.174 fractal_dimension_mean 2512.386	concave.points_mean 192381.200 radius_se alogas_se	symmetry_sean 11831.056 texture_se 194772.147	
perimeter_se 40376_836	area_se 1100444_173	amothesis del 41225 G02	
Compactness_se 473636.453	concavity_se 778242,166	concure.points_se 1574955_418	
kymmetry_sie 24678.711	fractal_dimension_we 463075.212	rwdius_worst 3511806.059	
texture.worst	perimeter_worst 617752,395	4767.645.776	
0530041450	compactness_worst 91238.174	concwv+ty_wormt 1705825.543	
amothmes_worst 56738.790		actal_dimension_worst	
smuothmess_worst 58738.790 concare.points_morst 561062.026	Symmetry_worst 1 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

Edit Code View Flore	Section Bullet Pelson Drollin	Tools Liele	- 0 )
	+ Addr	a series and a	Smarph
Cerce			ach ach
Station()' on how to c	(te & ce & packages in pak	Neart Torns	
ope 'demo(D' for some of help.start(D' for an UT ype 'qO' to putt E.	emos, 'help()' for an-line MC broaser interface to he	herip, an p:	
canter-real cavi"data	-147°)		
dis(cancer)			
cannerline-Hora model - glacdingoosis	Fant Ty-bit meet at cit take!	logit'', datascieseri	
gim.Fit algorithm d) gim.Fit recred proba (theory)car)	d mit converge milition numerically 0 or	Internet Will	
14518063.73% 4518063.73% 6121651.630	texture mean 140816.298 secotheses_mean 213030.000	perfineter.mene 1000.015.721 rdepartfrees.mene 1140.070	
concavity_mean 105593.174 Fractal_almension_sean	concave points_rean 192381.200 radius_se	symmetry_shiwn 11851.058 Cesture_se	
3513.136 perimeter.se 49276.836	610335.622 drea_se 1106444_172	356773.142 numerifyrets_se 41333.601	
2100pac1r4542_54 473535.453 59094279.54	concevity_se 7?8242,166 FrantaT_disension_se	contave.points_se 15/405.410 redits_worst	
24674.711 texture_putt1 873001.476	463075,232 gerieter_worst 617732,395	3311808.099 лема_могит 8707.045.770	
sauothness_worst 58738.790 concave_points_worst	compactness_wirst 91238.174 symmetry_wirst Fr	<pre><ul> <li><ul>     <li><ul>         <li><ul></ul></li></ul></li></ul></li></ul></pre>	
S61062:026 Tithrary(base) kapps(wodel)	10409,339	193961,530	
1.1. DEPENDENCE OF STREET			

Condition's number value is 619077723 which is above 15. This difference in values indicates multicollinearity problem.

#### 3. Condition Index



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.

RStedin							- 0	×
File Edit Code View Plats Session Build Debug 7	Finitia Tools Help							
Real and the second second	Addasi +						3.14	ni mani i
Planetti Sateri								-11(1)
E. Discontra II de la e							How THI HOURY	
W1 Decardy +							Street 25 Connect	Ling t
Constant Suffrage Street & December 1								-111
29 404.102119								
30 399.991329								1.0
31 1001.0400/0								10.0
- bestap	nacimater mint and man	inistinger main r	one hard been	concavity main	contrava policity paint		orea moltomenth fet	
T1.1 & 773165a-07 2 026900a-08 5 030156a-06	2.174501e-08 7.858511e-07	5.288301e-06	5 1810154-05	6.874284e-06	7.320370#-0	5.853066e-06	1.008763a-06	
T2 1 1 523472e-05 3 932052e-09 6 370618e-05	3.720149#-10 1.1821356-05	4.7580856-05	1.0458710-05	3. #80576e-04	3.801403e-04	1.034220e-04	2.376857e-05	
13.1 8.247095e-06 9.708480e-07 4.764155e-05	3.702015e-67.2.579327e-65	1.067121e-05	7.766950e-05	2.888275e-04	2:817070e-00	1.582800#-05	6.826674e-07	
14.7 5.853471e-07 1.062064e-07 E.817118e-05	1.304012e-07 7.835738e-06	7.736873e-06	9.337901e-05	8,106303e-05	1.475954#-04	7.521257#-06	5.023131e-07	
[5,1 1,758601e-05 2,624364e-06 7,247081e-04	2,496951e-06 1,930158e-04	1.343824e-04	2.0358964-03	1,749446e-05	4_050545a-04	1.573574e-04	1.197453e-05	
16,1 5.611717e-05 1.720983e-06 1.785916e-03	1.086983e-06 6.359330e-05	2.827995e-04	1.338575e-05	1.038023e-03	2.111051e-01	2.435753e-04	7.541371e-05	- 10
17.1 1.421381e-05 7.540807e-09 5.615256e-04	1.124270e-10 1.794863e-06	1.917364e-04	2:060652e-04	3:471788e-03	8.693680e-03	4.189071e-05	1.780796e-05	
[8,] 1.976882e-05 1.942888e-08 1.934420e-04	1,190795e-08 3.613138e-05	1.400138e-06	4.090628a-04	2.201888e-05	2.347636e-03	1.127101e-08	3.843228a-05	100
[9,] 1.207914e-05 1.890800e-08 1.104121e-04	2.411109e-06 2.247687e-04	4.531372e-04	5:454732e-04	4.2242376-04	1.641007e+03	1.045350e-05	5.480099e-05	
(10.7 A.473035e-08 2.375229e-07 3.551272e-04	3.350042e-07 2.722182e-04	1.320659e-06	3.305312e-05	9.848415e-03	1.938523e-0	1.639500e-06	2.787991e-00	
[11,] 2.342592e-05 9.577078e-07 4.978633e-04	1,110882e-06 3.963345e-04	1.261383e-03	3.841594e-04	8.611716e-04	2.284440e-03	1.798503e-01	3.127094e-04	1.1.4
[12,] 8.115133e-04 1.047644e-05 3.906645e-03	1.343765e-05 1.450199e-04	3.381914#-03	1.137460e-02	3.635046p-03	1.355026e-04	2.131877#-03	8.841319e-04	1.1
[1],] ].865888e-0? 1.236914e-0? 4,306892e-02	5,441728e-10 6,759761e-05	3.099374e-05	5.850412e-83	1.180390e-62	5.999519e-03	1 3.154732e-04	1.700100e-05	1.0
[14,] 1.419868e-04 9.830234e-06 6.208979e-05	1,104258e-05 3,837337e-05	1.740713e-03	1.390950e-03	1.564788e-03	2.854936e-03	1 7,274518e-01	2.493798e-05	
[15,] 8.066592e-04 2.793111e-05 1.669357e-03	3.122374e-05 1.870083e-03	2,924734s-04	9.032924#-03	1.545304e-03	3.460445#-04	1.827005e-02	4.085408a-04	
[16,] 4.022928e-04 5.170352m-06 8.744468e-04	3.227549e-06 6.870660e-06	2.79E458e-03	7.296732e-02	7.041698e-03	1.561171e-0.	3.701208e-02	5.4795328-04	
[37+] 3.767462e-04 6.522773e-07 3.997867e-04	2.862894e-06 1.657691e-03	2.533903e-02	2.362665e-02	8.432724e-03	3.498196e-0.	1.801771e-02	1.662352e-04	
[18] 1.779344e-04 4.069618e-05 1.713065e-01	4,343642e-05 5.029809e-03	4.635302e-03	1.066699#-03	7.268082e-02	L. 465393e-0.	A.70825/e-03	1.590928e-01	
[29,] 2.821167e-04 1.324383e-05 1.063713e-02	2,242219m-05.3.712368e-04	1.468550e-02	4.5197486-04	-2.0429026-02	4.9535598-03	8.477949#-03	4.664914e-03	
[20,] 4.140//2e-03 1.09918/e-0) 5.135289e-04	2.444008e-05 5.225403e-95	1.9158036-02	8,2//1/0e-02	4,600/28e-02	0.4424208-07	1.430389e-02	1,04080/e-02	
121.1 3.3106/8e-03.8.433096e-00 7.181125e-03	8,05(13/e-00-5,542/58e-95	4,187948e-W	8.724800e-02	1.1/0191e-0/	3.6/10/2/#-0.	1,7652/66-02	04/33/2-03	
[22, ] L.35838/4-03 3.200280s-05 4.41/784e-02	h.7185/2e-05 7.98//61e-05	1.4/99886-02	3.2190/20-03	3.9090838-01	2.5332230-10	1.3/43048-01	1.1385320-04	
(14 ] + Classe, 03 5 330804, 05 7 860434, 03	6 140003- 00 5.2888106-03	7.640347+.62	7.037504.02	1.0007020-01	3.20031198-04	6 273600- 02	3 025442- 05	
126 J 1 3 702404 05 3 2045404 05 5 2106424 01	3 TELEXE, NO 1 700334, A3	1 325671- 01	2 2520114 02	7. 3780734 67	5 232044 e.03	1.580755- 01	5 770746+ 03	
136 1 1 0104724 03 1 0002874 05 2 6888514 01	4 100178p-05 1 055543p-03	3,737780-01	1.0539370-03	1 1128250 02	1.100085a-01	2 330804-01	5.6506140-03	
[37] 1 x 206012a-02 1 640611a-03 3 221601a-03	2 376378n-fd 3 \$77707n-62	0.333(13e-02	1.238070a-03	3 154113-02	3.102821a-02	8 352853a-01	1.945683a-01	
128 1 7 980705e-01 2 886440e-04 2 895355e-03	2 302530e-04 1 148315e-02	0.763893a-03	2,275149a-01	3. 864830+.64	1.144503e-05	1 0096694-02	T. \$49878a.01	
179 1 3 400703e-07 1 087495e-07 9 602561e-04	6 554039e-03 7 403848e-01	3 2227826-04	8 0053318-03	1 1357350.07	1.394573#-03	2 6251614-05	4 1086976-01	
T30 T 8 885677#-02 1 #40708#-02 1 272639#-02	4.4710V0=-62 5.792510=-01	1.045900e-01	5 154311e-02	4.167885e-03	1.168633+-04	T 613819+-02	1.538U/4e-02	
[31.7 9.101913e-03.9.736573e-01.5.861733e-05	9,478888e-01 7,217100e-02	2:173449e-02	2,964350e-01	6.716995e-02	1.170140e-04	2.667463e-01	1.026853e-02	
radius se texture se perimeter se	area se seporteress se c	pepactness se con	cavity se concar	e paints se sy	mustry as fractal a	tinens im sa radius e	agent texture worst	
[1, ] 4.639117e-06 4.432415e-05 5.388969e-06	1.16/393e-03 4.263420e-05	2.394378e-05 2.7	99438e-03 2	2.291405e-05 J.1	12223e-05	1.467458e-05 1.335760	le-07 3.669378e-06	
[2,] 1.364898e-04 1.336559e-03 1.701150e-04	1.496404e-03 1.397966e-03	1.687406e-06 1.0	50453e-04 0	1.235532e-06.7.4	07763e-04 7	.798771e-04 2.436267	e-08 4.720149e-05	
[3,] 1.064301e-05 8.539957e-05 6.360532e-05	6.119387e-04 1.789819e-05	2.888747e-03 8.5	4467 he-03 - 3	1.513237e-04 3.7	63139e-06	.356751e-03 7.706132	be-05 1.500542e-05	
[4,] 7.036558s-04 4.509101s-03 7.581565e-04 1	2.788926a-01 1.857466e-01	4.870120e-D4 1.4	91703e-03 (	6.483306e-04 1.0	07758e-03	174719e-01 1.183320	le-06 3.044083e-05	
[5,] 4.964729e-04 4.316015e-03 % 303981e-04	1.95445te-01 0.517688e-03	6.402137e-04 4.4	57305e-02 4	1.437256e-03 7.9	24304e-03	753233e-05 1.223296	e-05 4.705482e-04	4
								-
						A 14 AL &	10:15 µµ	70
		-					18/9/2017	15

#### 4. Variance decomposition-proportion

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

1. Co. Co. Co. Co. Co. Co. Co. Co. Co. Co			Rotanant Ibarry
a Course lance			The second
a laste			Darts
LINEL.	M1 HC2 HCF	R(4 PIS	Beat 309 m. D
10.0001	0.[[###2244 0.23381/132 08.00#531249	9.001008962 - 0.01796534	No. 349 m. 5
arter penal	0.22751729 0.215181361 -0.004118270	L SCISCIUT - O. HAT I BAR	Wellach
-hept	-0.27089409 0.231076715 0.020689526	8.053433785 -0.00031223	Geel Large E.
Cheepis_amer	-0.13/38809 -0.580119023 -0.100291908 0.20038383 -0.551801838 -0.024091371	6. Eventski (J. 16) Sentiski Eventski (J.	MACK. LOUP M.
Arily sean	-0.25842008 -0.000105264 0.000723818	0.01012271 -0.0007511J	por sur LL
Here and a second second	-0.25685176 0.034767308 -0.025563541	0.0015531044 0.0017881021	orentation in .
etry_solar	-0.11819996 -0.787999777 -0.9817799800	E DATATION U. ANDREAS	
13.59	-0.20597878 0.101112052 0.208481307	8.83/941241 0.154656496	A 161 100
47.E.34	-0.00742803 -0.089079682 0.574E33803 -	-E. 3938/13/28 0. 2808/07/06	the res here
Later La	-0.2113/942 0.00002/234 0.30004382	E-05000048 0.12004010	1.0.0 10.0
STATUS OF	-0.01453141 -0.206450453 IL MORESBURY	0.044664183 0.212063678	a product ( Dalaman
en treppi, per	-0.1/109345 6.712/13896 0.154778/18	-0.027480981 - 47.229968138	promi
sunty an	10.18342740 -0.101201963 0.17046140	0.001248991-0.101990004 8.022800233 - 0.1015000	(stats) Documental
stry, so	-0.04249847 -0.583948080 0.288584292	0.0440/3351 0.212868/05	120.000
tall dimension are	-0.11/254837 -0.280000027 0.211593764 -0.212598001 0.110400028 -0.042596900	8. 0(1)/002/97 (0. 16/29/4.98 6. 0(1)/17.86 (0. 000408/00)	Principal
are series	-0.10849933 0.005467298 -0.042297828 -	-0.522607881, 0.062883000	Componen
hetsh_wirst	-0.23683908 D.229878428 -0.048548508	9.81380278# -0.007454151	Analysis
June Part of Contract	-0.1226/051 0.2299/1018 -0.00199/2418	4. 0.73004.98 0.30738002	Analysis
at trents, stors?	-0.10088588 -0.103092171 -0.296675625	-0.0911/20011 -0.17180010 <sup>4</sup>	And the second second
anity_Rorat	-U.12874713 -D.997964354 -U.173057395 -	-0.4*3011e0 -0.1ee514721	Description
sis.pornts_aneut	10.12290836 -0.181265361 -0.271112687	E SOLDERING - C.D. SJANOV	Perform a
tal allemater. sorit.	-0.13174394 -0.275339488 -0.212781313	-0.077031479 -0.064420031	arringel .
in ments	POS PCT P	CF PCB	AMOUNT OF FR
NTR. PERM	-U.0021786368 0.0115555382 -D.1596745	23 0.11260390	given bids mains
tetter, medal	0.0073084449 -0.1144770575 0.0085677	14 -4.222734211	and science the
	(0) HOLEST 480 - B. ONLES M275 - D. S MAC Set A. INTERNAL & LONGROUPS D. INTERNAL	40	of Taken is compare
automa jammi	-0.0141309488 0.0309188990 0.1512963	10 0. 10 V # # 10 V	and a second
pirity_lanati	-0.0093441800 -8:1071284434 0.8728272	85 0.040185000	Liteage
end benefit water	-0.0525499505 -0.1504822542 0.1523254	14 0, 1114/11300	and the second s

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:

Southe		Textesnant Haires
Content area, antrolit, emolithmess, worst compactness, worst compactness, worst contawn, perforts, journet wenetics, worst fractal_dimension_worst outpare, unan potame, unan potame, unan	-0.03828995 0.231359525 0.237362466 2.3896638-61 -0.04756678 0.022607466 -0.060633588 -1.532488-01 -0.04756678 0.022607466 -0.060633548 -1.532488-01 -0.1377034 -0.206633781 -0.16205348 -4.803123-02 -0.1377034 -0.206633781 -0.126203460 -1.7604986-02 -0.417034 -0.15557400 -0.25603488 -2.4570478-01 -0.4422952 0.08384296 0.022700444 -0.0036218-03 -74031388 -0.04275482 -0.035960388 -2.3567148-07 -220 FC30 -1.031934-03 -0.004278481 -3.318474-06.604290493	A Determine of the second seco
Pression modelfreats.mean compactives.mean compactives.mean merane.printl_mean yearity_mean metals_dimention(compactive) contains_te periodefic.co metals_dimention_an measure	$\begin{array}{rcl} -4,275948-01 & -0.0139475482 \\ -3.436657-03 & -0.0044764767 \\ -3.436657-03 & -0.00447647862 \\ -4.030479-02 & 0.0046743863 \\ -4.030479-03 & -0.01390681 \\ -4.286496-03 & -0.00127803794 \\ -7.086852-03 & -0.0021803794 \\ -7.3864320-03 & -0.0047303794 \\ -7.8864320-03 & -0.0047303794 \\ -7.886962-03 & -0.0047303794 \\ -7.8898-03 & -0.0047303794 \\ -7.899854-03 & -0.0047303794 \\ -7.899854-03 & -0.0047303794 \\ -7.899854-03 & -0.004873097 \\ -7.899854-03 & -0.004873097 \\ -7.899854-03 & -0.004892566 \\ -7.899854-03 & -0.004892566 \\ -7.8997140-03 & -0.0021978386 \\ -5.4997194-03 & -0.002199386 \\ -5.4997194-03 & -0.002199386 \\ -5.4997194-03 & -0.002199386 \\ -5.4997194-03 & -0.002199386 \\ -5.4997194-03 & -0.002199386 \\ -5.4997194-03 & -0.002199386 \\ -5.4997194-03 & -0.002199386 \\ -5.4997194-03 & -0.002199386 \\ -5.4997194-03 & -0.002199386 \\ -5.4997194-03 & -0.002199386 \\ -5.4997194-03 & -0.002199386 \\ -5.4997194-03 & -0.002199386 \\ -5.4997194-03 & -0.002199386 \\ -5.4997194-03 & -0.002199386 \\ -5.4997194-03 & -0.002199386 \\ -5.4997194-03 & -0.002199386 \\ -5.4997194-03 & -0.002199386 \\ -5.4997194-03 & -0.002199386 \\ -5.499914-03 & -0.002199386 \\ -5.499914-03 & -0.002199386 \\ -5.499914-03 & -0.002199386 \\ -5.499914-03 & -0.002199386 \\ -5.499914-03 & -0.002199386 \\ -5.499914-03 & -0.002199386 \\ -5.499914-03 & -0.002199386 \\ -5.499914-03 & -0.00219946 \\ -5.499914-03 & -0.002199494 \\ -5.499914-03 & -$	Ten Meter Meter Procession Procession Principal Component Analysis
odies_morst cotarschat primetorentit mea_unit coopditiesmorst coopditiesmorst coopditiesmorst constructentit presentinternit interiainternit interia_interniter interia_interniter	-6,372949-00 -0.1394439563 1.7733440-00 -0.01994395360 2.2931900-00 -0.0199439536 4.4493994-01 -0.0199429536 5.381400-03 -0.0044827778 5.3948046-06 -0.0124452778 5.3948046-06 -0.01245278 5.3948046-06 -0.01245278 5.30494378-02 -0.000482188 5.3240956-02 -0.0002884438 1.14477666-02 -0.001804435 1.1447666-02 -0.001804255	Description Pretorm a proceed components anayles on the generative an object installe as a religion of deep proceed

# 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.

🕒 Ršindio	-	a x
File Edit Cade Wew Plats Session Ruld Debug Profile Taols Help		
2 · · · · · · · · · · · · · · · · · · ·		Breach Hour +
Plants lanes		
the second seco	fSans +/ = 📑 6	Compart 2000
ET This lands a	There a	ist of their present -
Construction     0.00000000000000000000000000000000000	en or of the of	n 201 n. n 2

#### Model after PCA



11

Institutional Repository - Library & Information Centre - University of Thessaly 13/06/2024 07:17:16 EEST - 18.217.112.192

#### Model before PCA

and the second s									
data = rancer.p									-1
toefficiants: (Intercept) -0.2057 -1	+c1 rc2 4384 2.0197 -0	ec3 ec4	PC5 1,0912	PC6	PC7 0.4161	-1,3567	PC9 7.1282	Pc10 -1.0747	
auf) Deviance: testimal Deviance: - cancer-read.csv( - cancer-fac-tag) - cancer-fac-t	733,4 64.11 ALC: #6.11 Feata.csv7 ALC: #6.11 mendia con concerns probabilities namerically () = elegenets family = 1	nes"logit"), datamian or Focus ed inenia)()tek = "lagi	ero.						
data = tanter)	- magnetic - of carriers - o	and a state of the							
[Liter:	ept) radius_m	an texta	re_sean	perimeter	(.ietar)				
	2+00 F-92784	49 E-	Source and	concertly	r, eean				
-7.881 area_	man smoothness_m	as compactive							
-7, 881 area, -1, 301 concave.poliits	mean smoothrmss.ms e+05 -1.525e- neam symmetry.ms	at compactive 08 -f. An fractal_misensi	428e+06 on_peam	1.0	42e+06 1e5.5e				
-7, eel area -1, 301 concave.points. -1, 716 testur	naam sansthrass.ge e-05 -1.525e- neam symetry.ge e-07 4.045e- n_Ae perimeter.	an compactive 08 -6. an fractel_ansensi 07 -4.	428e+06 on_pean 233e+07 er#4_se	1.04 raf 3.32 secother	42e+06 165.5e 28e+07 85=_3#				
-7, eeu area, -1, 301 concave, polecs, -1, 716 testue 4, 368 compacteo	Maximum     Missisterman     Missisterman       ex00     -1.525e     -1.525e       ex01     symmetry.me     -1.615e       ex02     4.045e     -1.761e       ex06     1.761e     -1.761e       ex06     0.00000000000000000000000000000000000	an compactne 08 -f. An fractal_misensi 07 -4, se 06 -E. Ne corravo.ps	428e=06 on_mtan 233e=07 ant4_se 393#=05 ints_se	1.04 rad 3.32 inscorting 7.40 symmet	42e+06 Tec.3c 28e+07 e1=_30 82e+08 try_am				
7, 881 arms, -1, 301 conclave.goints, -1, 716 texture 4, 368 compactnes -1, 773 fractal_shimersio	start     stmothensis     stmothensis <th< td=""><td>an compactive 08</td><td>428e=06 on_mean 233e=00 ansa_se 293e=05 ints_se 260e=09 e_wors1</td><td>1.04 raf 3.12 inscorting 7.40 symmetry 2.87 per/instor/</td><td>42e+06 145.5e 28e+07 e15.5e 92e+08 92e+08 90e+08 90e+08</td><td></td><td></td><td></td><td></td></th<>	an compactive 08	428e=06 on_mean 233e=00 ansa_se 293e=05 ints_se 260e=09 e_wors1	1.04 raf 3.12 inscorting 7.40 symmetry 2.87 per/instor/	42e+06 145.5e 28e+07 e15.5e 92e+08 92e+08 90e+08 90e+08				
-7,881 area, -1.301 concave.points_ -1.710 treature 4.308 compactnes -1.773 fractal_dimension 1.512	maan samothemaa ge ee05 1.525e seam symmetry se ee07 4.045e maar perimeter, ee06 1.752e seam someter, se of 1.526e tiss ee and 1.526e ee09 -5.130e	at compatible 03	428e=06 on_mean 233e=07 ante_se 393a=05 ints_se 260e=09 e_worst 632e=05	1.04 riaf 3.13 uncothus 7.40 symmet 2.81 perfunctor, -3.33 reservation	42e+06 1e5_5e 28e+07 #82_0# 00m+08 50e+08 _movst 38e+05 				
-7, 661 area, -1, 301 concave, polyec, -1, 710 trastau -0, 308 compactrass -1, 773 fractal_atimestio 1, 512 orea, a -8, 950	naam ea05 1.525e. neum symmetry se ea07 4.045e make perimeter, action of the symmetry se ea06 1.752e. tase concurring, ea06 1.529e. tase concurring, ea06 - 1.529e. ea07 - 5.130e. ea07 - 2.151e.	at comparise 03 - f. an fractal_fismsi 05 - 4. 06 - 4. 06 - 4. 06 - 2. 06 - 3. 08 - 3. 08 - 3. 01 - 1. 09 - 4. 00 - 4	428=-06 orr64ii 233=407 arit4_se 233=05 rint1_se 260e-09 fworst 632=05	1.0 raf 3.17 useotha 7.40 2.39 pertmeter -3.5 roscavity, -3.0	42e+06 1e5_5e 28e+07 91s_3e 92s+08 try_ss 90e+08 _wo/St 30e+05 _wo/st 28e+07				
- 7, 660 arrea, - 1, 301 concinve, pointsc. - 1, 710 tratue - 4, 308 compact time - 4, 773 fr actal_diseasion - 1, 512 - arrea, - 4, 500 concinve, pointsc. - 4, 500 concinve, pointsc. - 1, 431 - 1, 431	maximum     same difference       e-05     1.525e       neam     symmetry are       e-07     4.049 are       e-06     1.752e       state     perimetery       e-06     1.752e       state     constructory       e-06     1.752e       e-06     1.752e       e-07     6.0401ms       e-08     1.32e       e-09     -6.130e       e-04     -2.152e       e-05     2.474e	at compactive 303 - f. an fractal_fisensi 306 - 4. 306 - 4. 306 - 5. 507 - 5. 507 - 5. 507 - 4. 507 - 5. 507 - 5. 5	428=06 on_melan 233=07 aridse 293=07 ints_se 260=05 f_sec555 f_sec5555 f_sec5555 f_sec5555 f_sec5555 f_sec5555 f_sec5555 f_sec5555 f_sec5555 f_sec5555 f_sec5555 f_sec5555 f_sec5555 f_sec55555 f_sec5555 f_sec5555 f_sec55555 f_sec55555 f_sec55555 f_sec555555 f_sec55555 f_sec55555 f_sec555555 f_sec555555 f_sec55555555 f_sec5555555555 f_sec555555555555555555555555555555555555	1.00 raf 3.33 iscottas 7.40 syssed 2.83 per/uncter -3.33 covcav(ty, -3.00	42e+06 NeLie 28e+07 PE-10 PE-10 PE-08 PE-05 NeT 28e+05 NeT 28e+07				
-7,601 area, -1.301 Concave, pointsc. -1.710 Statue -4.773 fractal_dimension -1.773 fractal_dimension -1.431 area.a concave.points.a -1.431 concave.pointsc.a -1.431	maximum     samedthemail, we       e-05     1.525c       neam     symmetry, we       e-07     4.049c       e-08     9.071000       main     symmetry, we       e-06     1.752s       s,se     0000000       s,se     00000000       e-06     1.52s       e-05     1.52s       e-06     1.52s       e-07     6.045s       e-08     -2.152s       e-05     -2.474s       2465     -2.474s       2465     -2.474s	at compactive fractal_alienci fractal_alienci at conceau.go do conceau.go do conceau.go do conceau.go to to to to fractal_alienci a at conceau.go to to to to to to to to to t	428e-06 on_mtar 233e-00 arms_se 293e-05 ints_se 260e-09 4_mors1 852e-05 file=05 fil	1.04 rad 3.3 secottos 7.40 system 2.8 per lateter -3.5 coscavity, -3.04	42e+06 Hat.de Read7 Mix.am Mix				

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)</pre>

#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)</pre>

percentVariance <- pca.cancer\$sd^2/sum(pca.cancer\$sd^2)\*100

percentVariance

library(caret)

trans <- preProcess(cancer.predictors,method = c("center", "scale", "pca"))</pre>

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"