International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2019): 7.583

Selection of the Number of Components to Keep in Order to Build High Quality Regression Model

Solovei Olga[1], Solovei Bohdan[2]

[1]Kyiv University of Civil Building and Architecture, Kyiv, Povitroflotsky Avenue, 31, 03680, Ukraine
soloveiolga2[at]gmail.com

[2]National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" Peremohy Ave, 37, Kyiv, 03056, Ukraine
bsolovei25[at]gmail.com

Abstract*: in the article is considered a question of how to identify the number of components to keep when using principal components analysis technique for dimension reduction. In paper is presented the impact of incorrectly chosen number on the quality of regression model; reasons why the number can be identified incorrectly using principal components analysis. The summary proposes the method for identification the number of components to keep which can be used together with principal components analysis technique for dimension reduction.*

Keywords: Principal component analysis (PCA), Singular Value Decomposition (SVD), eigenvalues, Kaiser rule.

1. Introduction

Principal component analysis (PCA) is the main technique in unsupervised machine learning that is used to reduce dimensions so that the low-dimensional representation retains some meaningful properties of the original data [1].

PCA can be performed in Python by calling function PCA from library sklearn.decomposition [2], however, the number of principal components is required as an input parameter – it defines the number of components to keep.

The impact of choosing incorrect number is obvious and can be illustrated with the standard diabetes dataset embedded in Python library sklearn.datasets through the following steps: 1) call PCA function with number of components to keep equal to 1; 2) transform training and test data into principal components; 3) build linear regression model; 4) evaluate r-squared score of built model. 5) repeat steps 1)-4) with number from 2 till 10.

Table 1: R^2 score of regression model of diabetes dataset

number of components to keep	R^2
1	0.33
2	0.35
3	0.36
4	0.48
5	0.48
6	0.46
7	0.47
8	0.47
9	0.47
10	0.47

From Table1 can be concluded the only call of PCA with number of components to keep equals to 4 gives the best r-squared score; the call of PCA with any other number causes less r-squared score of the built model.

Therefore, in order to build high quality regression model it is important to select the proper number of components to keep.

2. Goal

To propose the method for identification the number of components to keep which can be used with PCA technique for dimension reduction to ensure the high r-squared score for regression model.

3. Main part

Three most common methods for selecting the number of principal components are: 1. Kaiser rule; 2. Scree plot; 3. Proportion of variance explained.

Kaiser rule identifies the number of components to keep to be equal to the number of eigenvalues which are bigger or equal to 1 [5]. Scree plot method calculates the number to be equal to the number of eigenvalues which are above the "elbow" [4]. The proportion of variance explained specifies to select the component to be kept when the proportion of its explained variance is higher a certain threshold, e.g. 70% or 90% and the proportion of variation explained for the i - th component is defined to be the eigenvalue for that component divided by the sum of the eigenvalues.

So all the mentioned methods are based on eigenvalues analysis which makes eigenvalues' calculation approach are critical for correct identification of the number of components to keep.

The eigenvalues λ in PCA technique are calculated by solving equation $\left|A_{\text{cov}} - \lambda I\right| = 0$, where A_{cov} is a covariance matrix of transposed centered matrix of independent features Ac.

Dimension reduction can also be performed using Singular Value Decomposition (SVD) technique by solving equation $\left|Ac^{'} \cdot A - \lambda I\right| = 0$, where A is a matrix of independent features and $Ac^{'}$ - transposed centered matrix of A [3].

Kaiser rule; scree plot and proportion of variance explained methods can be applied on eigenvalues received either from PCA or SVD techniques and should show identical results.

To verify the assumption, we perform eigenvalues' calculation using PCA and SVD methods first on a small dataset and then repeat the calculation on dataset of the bigger size.

Example 1.1. Apply SVD technique to identify the number of components to keep of the 2×3 matrix

$$A=\begin{bmatrix}1 & 2 & -3\\ 2 & -1 & 1\end{bmatrix}$$

Centered matrix A

$$Ac=\begin{bmatrix}-0.5 & 1.5 & -2\\ 0.5 & -1.5 & 2\end{bmatrix}$$

Form product:

$$Ac^{'}Ac=\begin{bmatrix}0.5 & -1.5 & 2\\ -1.5 & 4.5 & -6\\ 2 & -6 & 8\end{bmatrix}$$

Identify the eigenvalues of $Ac^{'}Ac$ matrix by resolving equation

$$\begin{vmatrix}0.5-\lambda & -1.5 & 2\\ -1.5 & 4.5-\lambda & -6\\ 2 & -6 & 8-\lambda\end{vmatrix}=0\Rightarrow$$

Obtain zeros in 2nd row of $Ac^{'}Ac$ matrix

$$\begin{vmatrix}5-\lambda & -1.5 & 2\\ 0 & -\lambda & 0\\ 20 & -6 & 8-\lambda\end{vmatrix}=0$$

Calculate determinate along 2nd row

$$-\lambda\begin{vmatrix}5-\lambda & 2\\ 20 & 8-\lambda\end{vmatrix}=0\Rightarrow-\lambda(40-13\lambda+\lambda^2-40)=-\lambda(\lambda^2-13\lambda)=0;$$

Therefore, the sorted eigenvalues are $\lambda_1=13$; $\lambda_2=\lambda_3=0$. Applying Kaiser's rule, we receive the number of components to keep is equal to 1.

Example 1.2. Apply PCA technique to identify the number of principle components of the 2×3 matrix A from example 1.1

Identify covariation matrix for transposed centered matrix $Ac^{'}$

$$A\text{cov}=\begin{bmatrix}0.5 & -1.5 & 2\\ -1.5 & 4.5 & -6\\ 2 & -6 & 8\end{bmatrix}$$

The covariation matrix $A\text{cov}$ is equal to product matrix $Ac^{'}Ac$ received in example 1.1 so eigenvalues are also equal and Kaiser's rule will show the same result as obtained in example 1.1.

Now we repeat the calculation from examples 1.1 for standard diabetes dataset from python library sklearn.datasets. When we load data then we received 224×10 matrix with independent features. Appling SVD technique we received eigenvalues, sorted in decreasing order:

$\lambda_1=4;\lambda_2=1.5;\lambda_3=1.2;\lambda_4=1;\lambda_5=0.7;\lambda_6=0.6;\lambda_7=0.5;\lambda_8=0.4;$
$\lambda_9=0.1;\lambda_{10}=0$

and according to Kaiser's rule - the number of components to keep is equal to 4 which corresponds to best r-squared scope from Table1.

Repeating the calculation from example 1.2 for diabetes dataset –we received covariation matrix which shows almost no correlation between features (figure 1).

$A\text{cov}=$

0.0023	0.0004	0.0004	0.0008	0.0006	0.0005	-0.0002	0.0005	0.0006	0.0007
0.0004	0.0023	0.0002	0.0005	0.0001	0.0003	−0.0009	0.0008	0.0003	0.0005
0.0004	0.0002	0.0023	0.0009	0.0006	0.0006	-0.0008	0.0009	0.001	0.0009
0.0008	0.0005	0.0009	0.0023	0.0005	0.0004	-0.0004	0.0006	0.0009	0.0009
0.0006	0.0001	0.0006	0.0005	0.0023	0.002	0.0001	0.0012	0.0012	0.0007
0.0005	0.0003	0.0006	0.0004	0.002	0.0023	-0.0004	0.0015	0.0007	0.0007
-0.0002	-0.0009	-0.0008	-0.0004	0.0001	-0.0004	0.0023	−0.0017	-0.0009	-0.0006
0.0005	0.0008	0.0009	0.0006	0.0012	0.0015	-0.0017	-0.0023	0.0014	0.0009
0.0006	0.0003	0.001	0.0009	0.0012	0.0007	-0.0009	0.0014	0.0023	0.0011
0.0007	0.0005	0.0009	0.0009	0.0007	0.0007	-0.0006	0.0009	0.0011	0.0023

Figure 1: Covariation matrix of diabetes dataset

Solving equation $|A_{\text{cov}}-\lambda I|=0$ for covariance matrix from figure1 gives almost all eigenvalues are closed to zero:
$\lambda_1=0.01;\lambda_2=\lambda_3=\lambda_4=\lambda_5=\lambda_6=\lambda_7=\lambda_8=\lambda_9=\lambda_{10}=0$.

According to Kaiser's rule the number of components to keep is equal to 0 – which mean no principal components exist in dataset and contradicts with the results obtained applying SVD techniques and results from Table 1.

4. Conclusion

Applying PCA technique for dataset with low covariation between independent features result that eigenvalues are closed to zeros and number of components to keep according to Kaiser's rule is zero which is not correct.

To identify correct number of components to keep when using PCA technique it is required first to calculate eigenvalues of matrix $Ac^{'}Ac$ then apply one of the rules 1-3 to identify the number of components to keep. Dataset transformed in that number of principal components will be well prepared for high quality regression model to be built.

References

[1] Westfall,P.H., Arias, A.L., and Fulton, L.V. (2017). Teaching Principal Components Using Correlations, Multivariate Behavioral Research, 52, 648-660.
[2] Python: Advanced Guide to Artificial Intelligence by Rajalingappaa Shanmugamani; Giuseppe Bonaccorso; Armando Fandango Published by Packt Publishing, 2018
[3] Linear Algebra and Matrix Analysis for Statistics by Sudipto Banerjee; Anindya Roy Published by Chapman and Hall/CRC, 2014

[4] Cattell, R. B. (1966). The Scree Plot Test for the Number of Factors. Multivariate Behavioral Research, 1, 140-161.
[5] Kaiser, H. F. (1960). The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20, 141-151.

Author Profile

Solovei Olga – received PhD in Computer Science in Kyiv University of Civil Building and Architecture in 2013. Takes position as Senior business analyst at Luxoft Ukraine and professor assistant of the department of applied mathematics in Kyiv University of Civil Building and Architecture.

Solovei Bohdan – Student of the 1st year of magistracy National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute