Principal Component Analysis– Performance Assessment

Steven Schindler

Data Mining II – D212

Kamara, Kesselly; PhD

College of I.T., Western Governors University

Table of Contents

Part I: Research Question

A1.

What is the optimal number of dimensions(principal components) needed for the telecommunications dataset?

A2.

There are 13 continuous variables in the data set the goal of this analysis is to reduce the number of variables for the sake of simplicity at the cost of accuracy.

Part II: Method Justification

B1.

PCA reduces the dimensions of a data set in order to find the principal components. The magnitude of the covariances indicate the strength of the correlation. The expected outcomes of PCA are dimension reduction and feature selection. The principal components with high eigenvalues indicate the feature selection. Then using dimension reduction, we can reduce the initial number of variables to the number of chosen principal components.(DataCamp. (n.d.))

The covariance of the initial variables are then used to calculate the principal components. A covariant matrix is created using these covariances then getting the eigenvalues from this matrix it is then used to get the principal components. These components are new variables that are linear combinations of the old variables.(Builtin. (n.d.))

B2.

One assumption of PCA is linearity, it assumes a linear relationship between variables.(Mueller, F. (n.d.).)

Part III: Data Preparation

C1.

The thirteen continuous variables are 'Lat', 'Lng', 'Population', 'Children', 'Age', 'Income', 'Outage_sec_perweek', 'Email', 'Contacts', 'Yearly_equip_failure', 'Tenure', 'MonthlyCharge', and 'Bandwidth_GB_Year'.

C2.

```
df = pd.read_csv('churn_clean.csv')
numeric_list = list(df.select_dtypes(include = {'int64','float64'}))
del numeric_list[:2]
del numeric_list[13:]
pca_df = df[numeric_list]
scaler = StandardScaler()
pca_norm = scaler.fit_transform(pca_df)
pd.DataFrame(pca_norm).to_csv('standarized_churn.csv', index=False)
```

Above is the code used to Standardize the data set. The new file is called standarized_churn.csv.

Part IV: Analysis

D1.

	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9	PC 10	PC 11	PC 12	PC 13
Lat	-0.023161	0.007911	-0.001230	0.014244	0.001860	0.004185	0.005811	-0.020020	0.004283	0.017665	0.705211	0.040456	0.706719

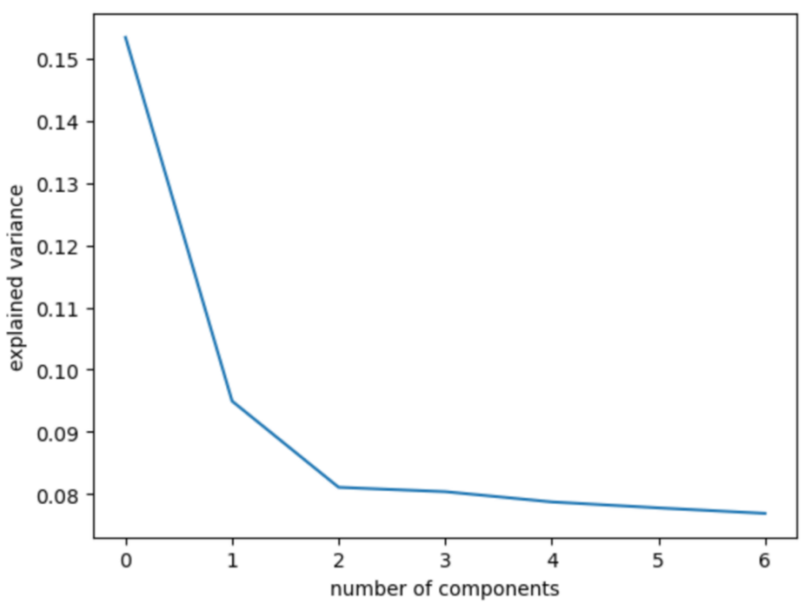
	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9	PC 10	PC 11	PC 12	PC 13
Lng	-0.714010	0.180879	0.653439	-0.014267	0.052795	-0.054602	0.009174	0.152355	0.031043	-0.007070	-0.008913	-0.004500	-0.010435
Population	-0.031715	-0.285753	0.151916	0.447882	-0.443537	0.195742	-0.249550	-0.092711	-0.447906	0.153686	0.006569	-0.404228	0.008289
Children	0.109414	-0.736871	0.322012	-0.464670	0.227235	-0.041772	-0.126214	-0.144998	0.108875	0.063449	0.026652	-0.136041	-0.002713
Age	-0.094872	0.344620	-0.119517	-0.107498	0.436759	0.312779	-0.455981	-0.353186	0.011245	0.420468	0.009197	-0.218356	-0.021522
Income	-0.030887	-0.087695	0.098791	0.130597	-0.096321	0.100371	0.597523	-0.403463	0.082442	0.592380	-0.036725	0.257205	-0.012558
Outage_sec_perweek	-0.010719	-0.052349	0.053682	0.034812	-0.188399	0.773549	0.051915	0.003835	0.519791	-0.290766	-0.002190	-0.041495	0.003902
Email	-0.020375	-0.086499	0.079161	-0.065531	0.093484	0.335467	-0.184658	-0.125375	-0.510974	-0.194665	-0.038433	0.714123	0.002926
Contacts	0.090273	-0.172285	-0.027392	0.192459	0.342892	0.246663	0.057056	0.760622	-0.052695	0.397088	0.003806	0.060669	0.002798
Yearly_equip_failure	0.018619	-0.151301	0.055304	0.437471	-0.083596	-0.275852	-0.515406	-0.052146	0.494601	0.143419	-0.037339	0.405280	0.005673
Tenure	0.053958	-0.112280	0.100818	0.565626	0.614892	-0.033742	0.223304	-0.247985	-0.028194	-0.376943	0.005491	-0.144376	-0.006102
MonthlyCharge	0.674376	0.375138	0.631729	-0.011794	-0.037729	0.006645	-0.034155	0.027357	-0.011878	0.038880	0.010393	-0.006021	0.009429
Bandwidth_GB_Year	0.001077	0.000788	-0.000070	-0.021597	0.022360	-0.000941	0.000271	0.000274	-0.000947	-0.000083	-0.705254	-0.0457	0.706791

D2.

Using the Kaiser criterion, we eliminate PCs with an eigenvalue less than 1 all the eigenvalues are shown in the screen shot below.

	Eigenvalues
PC 1	1.99
PC 2	1.23
PC 3	1.05
PC 4	1.04
PC 5	1.02
PC 6	1.01
PC 7	1.00
PC 8	0.99
PC 9	0.98
PC 10	0.96
PC 11	0.96
PC 12	0.74
PC 13	0.01

Using this table, we see that we can reduce the dimensions to 7 eliminating PC8 – PC13. Below is the resulting scree plot:



D3.

Each of the principal components has a variance that is listed in the table below:

	Variance
PC 1	15.34
PC 2	9.49
PC 3	8.11
PC 4	8.04
PC 5	7.87
PC 6	7.77
PC 7	7.69

D4.

The total variance being the sum of each variance in D3 is 64.31.

D5.

The results of the PCA show we can group the 13 variables into 7 more manageable groups. The table below highlights which variables have the strongest contributions to the principal component. Since the highlighted numbers mean they have a strong linear combination they are also correlated, and these highlighted numbers also have a strong impact on their respective PC. (BRUCE, P. A. (2020)) For example, Lng and MonthlyCharge have a strong impact on PC 1, so they also have a strong linear combination and are correlated. The optimal number of dimensions for this data set is 7.

	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7
Lat	-0.023161	0.007911	-0.001230	0.014244	0.001860	0.004185	0.005811
Lng	-0.714010	0.180879	0.653439	-0.014267	0.052795	-0.054602	0.009174
Population	-0.031715	-0.285753	0.151916	0.447882	-0.443537	0.195742	-0.249550
Children	0.109414	-0.736871	0.322012	-0.464670	0.227235	-0.041772	-0.126214
Age	-0.094872	0.344620	-0.119517	-0.107498	0.436759	0.312779	-0.455981
Income	-0.030887	-0.087695	0.098791	0.130597	-0.096321	0.100371	0.597523
Outage_sec_perweek	-0.010719	-0.052349	0.053682	0.034812	-0.188399	0.773549	0.051915
Email	-0.020375	-0.086499	0.079161	-0.065531	0.093484	0.335467	-0.184658
Contacts	0.090273	-0.172285	-0.027392	0.192459	0.342892	0.246663	0.057056
Yearly_equip_failure	0.018619	-0.151301	0.055304	0.437471	-0.083596	-0.275852	-0.515406
Tenure	0.053958	-0.112280	0.100818	0.565626	0.614892	-0.033742	0.223304
MonthlyCharge	0.674376	0.375138	0.631729	-0.011794	-0.037729	0.006645	-0.034155
Bandwidth_GB_Year	0.001077	0.000788	-0.000070	-0.021597	0.022360	-0.000941	0.000271

References:

Mueller, F. (n.d.). Lec 17 - Principal Component Analysis. Retrieved May 21, 2023, from https://www3.cs.stonybrook.edu/~mueller/teaching/cse564/Lec%2017%20-%20Principal%20Component%20Analysis.pdf

DataCamp. (n.d.). Principal Component Analysis in Python. In DataCamp. Retrieved May 21, 2023, from https://www.datacamp.com/tutorial/principal-component-analysis-in-python

Builtin. (n.d.). Step-by-Step Explanation of Principal Component Analysis. In Builtin. Retrieved May 24, 2023, from https://builtin.com/data-science/step-step-explanation-principal-component-analysis

BRUCE, P. A. (2020). Practical statistics for data scientists. 50+ essential concepts using r and python.
O'Reilly Media, Incorporated. WGU Library.