Principal Component Analysis with SVD

Introduction

Recent studies of The Tropics region across a broad range of social and economic indicators (https://www.jcu.edu.au/state-of-the-tropics/publications/2014/2014-report/State-of-the-Tropics-2014-Full-Report.pdf) indicate that it becomes prominent as a critical global region with a distinctive set of opportunities. It covers only 40% of the world's surface area and it is a home to 40% of the world's adult population and 55% of the world's children population under the age of five. The Tropic's economy is growing 20% faster than the Rest of the World but still more than two-thirds of the world's poorest people live in the Tropics. Significant poverty reduction is observed in South East Asia and Central America. Persistently with the higher levels of poverty, more people experience undernourishment, even though the undernourishment rate in the Tropics has declined by one-third over the past two decades. The urbanization rate has increased from 31% to 45% in 2010. Life expectancy has also increased across all regions of the Tropics and there is a significant decrease in maternal and child mortality rates since 1950. Mean years of schooling of adults has almost doubled in the Tropics since 1980, adult literacy rates have increased faster in the Tropics than the Rest of the World and youth literacy rates have increased in all regions of the Tropics except in Oceania since 1990, but still steadily lower than the Rest of the World.

In current study, the social and economic indicators listed in (Table 1) are subjected to Principal Component Analysis (PCA) with SVD (Singular Value Decomposition). Principal component analysis as a variable reduction procedure is appropriate when we have obtained data on a number of observed variables and believe that there is some redundancy among the variables. Therefore, we wish to develop a smaller number of artificial variables (called principal components) that will account for most of the variance in the observed variables that may be used as criterion variables in subsequent analyses. The first component extracted accounts for a maximal amount of total variance in the observed variables. The second component will have two important features: a) it accounts for a maximal amount of variance in the data set that was not accounted for by the first component and it will be correlated with some of the variables that did not display strong correlations with the first component; b) second feature is that it will be uncorrelated with the first component.

PCA is connected to SVD that is a highly robust method because of its ability to decompose any matrix into $U\Sigma V^T$ resulting with two orthogonal matrices U and V and diagonal Σ. Nevertheless, in this study, SVD is applied on the matrix of observations by decomposing the data matrix into three factors that contain the eigenvectors and eigenvalues utilized in the PCA method. The study demonstrates the interpretative power of PCA with SVD by carrying out an exploratory data analysis with final goal to determine relationships that exist between and within the data structure of a statistical data from The Tropics.

Data

The dataset of interest, called SotTCombined2010, is retrieved from the State of the Tropics website located at the following URL: https://www.jcu.edu.au/state-of-the-tropics. It is available as spreadsheet in MS Excel® (.xlsx) extracted from various existing data sources for a period of 2010, such as:
FAOdatabase (http://www.fao.org/faostat/en/#data); KILM database (http://kilm.ilo.org/kilmnet/); Barro and Lee Educational Attainment dataset (http://www.barrolee.com/); Povcal Net, World Bank database (http://iresearch.worldbank.org/PovcalNet/povOnDemand.aspx); WHO Global TB database (http://www.who.int/tb/country/global_tb_database) and etc.
***Size of the dataset**:* The size of the dataset is 109 rows with 14 variables as shown in Table 1.

	Name	Data type	Description
1.	Country	Character	Name of the countries involved in the research
2.	Life Expectancy (years)	Numeric	Calculated average of life expectancy in years
3.	Poverty under $1.25 per day	Numeric	Extreme poverty measure for developing nations
4.	Population under 15	Numeric	Total population under 15 years of age in each country
5.	Adult literacy (% above 15)	Numeric	Calculated adult literacy in percentage above 15 years of age
6.	Mean years of schooling	Numeric	Estimated mean years of schooling by nation
7.	Unemployment (%)	Numeric	Estimated unemployment rate in percentage by nation
8.	Youth literacy (% age 15-24)	Numeric	Estimated percentage of youth literacy between 15-24 years old
9.	Tuberculosis (cases)	Numeric	Number of tuberculosis cases
10.	Under 5 mortality	Numeric	Calculated mortality under 5 years of age per 1000 births
11.	Poverty under $2 per day	Numeric	Estimated number of people living under poverty line ($2 per day)
12.	Undernourishment	Numeric	Estimated number of people at risk of undernourishment
13.	Urban population (%)	Numeric	Calculated urban population as percentage of total population
14	Area of agricultural land (??)	Numeric	Area of agricultural land by nation

Table 1: Data description of the SotTCombined2010 dataset consisting of one categorical/character variable and thirteen numeric/continuous variables.

Importing the dataset

```
% Import raw data from excel file
Xshift = xlsread('SotTCombined2010.xlsx');
```

We call raw data matrix Xshift, a rectangular matrix with size (109x13), imported from excel file (SotTCombined2010.xlsx) with *xlsread()* code into MatLab environment where number of columns/measured variables is presented by n, (n=13) and the number of row or "cases" which have had all variables measured is presented by m, (m=109). The code, *xlsread(filename)* reads the first worksheet in the MS Excel® spreadsheet workbook named filename and returns only the numeric data in a matrix. As a result, the categorical variable "Country" is omitted.

Pre-processing the dataset

Excluding countries with missing values

```
% Remove missing values using rmmissing() function
Xshiftrm = rmmissing(Xshift);
```

Using the ismissing(Xshift), sum(isnan(Xshift)), sum(any(isnan(Xshift)) commands present in the LiveScript, we determined exactly 187 missing values or "NaN" distributed in 9 columns and 53 rows that are removed with *rmmissing ()* code resulting into reduced matrix size (from 109 to 53 rows) of a newly created matrix called Xshiftrm (56x13).

Creating, centering and scaling matrix of raw data for PCA

PCA only deals with a square matrix containing only numeric data, but SVD can deal with a rectangular matrix. For the purposes of this assessment, we will apply an SVD to perform a PCA. Before applying the SVD, we need to pre-process the data table by mean-centering the data. Primarily, we compute the mean row vector and the variance row vector for all data vectors:

```
% Compute means and variance of each column
colmeans = mean(Xshiftrm,1)
colvars = var(Xshiftrm,1)
```

Therefore, subtracting the mean row vector from all the data vectors and dividing by the square root of the variance vector to avoid bias caused by large variations among variables resulted into creation of a new matrix called X-scaled matrix (56x13).

```
% Compute the X-scaled matrix from the raw data
X_scaled = (Xshiftrm - repmat(colmeans, size(Xshiftrm,1), 1)) ./ repmat(sqrt(colvars),
size(Xshiftrm,1), 1)
```

Total Variance – Summary

X-scaled matrix (56x13) contains variables that are standardized so that each variable has a mean of zero and a variance of one. The total variance in the data set is the sum of variances of all observed variables. Now, because each variable has a variance of one, it contributes only one unit of variance to the total variance in the data set. As a result, the total variance in PCA will always be equal to the number of variables being analysed. For example, we have 13 variables available in the current study, the total variance will always equals to 13. The PCA components that are going to be created will always partition total variance.

Methods and Analysis

The application environment for the project is as following: Matlab Version R2019b Update 1(9.7.0 2126025).

Perform PCA using Singular Value Decomposition (SVD)

The most essential step is performing SVD. As one of the most important linear algebra principles, SVD is utilized to calculate a lower dimensional space of PCA. The aim of the SVD method is to diagonalize the data matrix X_scaled (56x13) into three matrices: U, S and V. In general, U(56x56) matrix represents left singular vectors, S(56x13) is a diagonal matrix comprised of singular values that are sorted from highest-to-lowest and must satisfy the rule that the first singular value is the largest and each succeeding singular value are smaller:

σ1 > σ2> σ3 > σ4 > σ5 >………> σn

And V (13x13) stands for right singular vectors, the non-zero eigenvectors or the principal components of the PCA space. The left and right singular matrices, i.e. U and V, are orthogonal matrices. From the calculated matrices U, S and V, we can notice that the first singular value of S matrix was more than the second one, thus, the first column of V represents the first principal component and so on.

```
% Compute Singular Value Decomposition that is equivalent to X_scaled = U*S*V'
[U,S,V] = svd(X_scaled)
```

```
U = 56×56
  -0.1557   0.0538   0.1721  -0.2918  -0.0234  -0.0794   0.0126 ⋯
   0.0361   0.0625  -0.2533   0.0890   0.0461  -0.1820  -0.2519
  -0.2625   0.0646  -0.0247   0.3264   0.1701   0.0890   0.1420

S = 56×13
  18.7152        0        0        0        0        0        0 ⋯
        0  11.4672        0        0        0        0        0
        0        0   8.2715        0        0        0        0

V = 13×13
   0.3421  -0.0551   0.2105  -0.0248  -0.1184   0.2813   0.5508 ⋯
  -0.3579  -0.0097  -0.0040   0.2234   0.2099  -0.1617   0.2667
  -0.0095  -0.6408   0.0490   0.0379  -0.1113  -0.1692   0.0240
```

Identifying relationships between the variables described by the 1st principal component vector

We can notice (from Table 2, Fig 1 below) that the first eigenvector as the strongest PCA component has positive and negative elements.

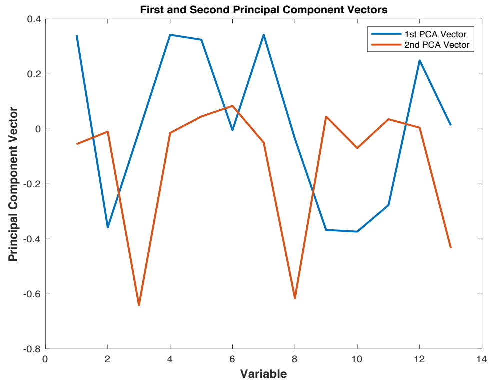


Fig 1: Min and max loading points of 1st and 2nd PCA vectors

Variable	Variable name	1st PCA vector	2nd PCA vector
v_1	Life Expectancy (years)	0.3421	-0.0551
v_2	Poverty under \$1.25 per day	-0.3579	-0.0097
v_3	Population under 15	-0.0095	-0.6408
v_4	Adult literacy (% above 15)	0.3425	-0.0146
v_5	Mean years of schooling	0.3246	0.0453
v_6	Unemployment (%)	-0.0042	0.0839
v_7	Youth literacy (% age 15-24)	0.3427	-0.0494
v_8	Tuberculosis (cases)	-0.0354	-0.6156
v_9	Under 5 mortality	-0.3672	0.0451
v_{10}	Poverty under \$2 per day	-0.3734	-0.0694
v_{11}	Undernourishment	-0.2771	0.0355
v_{12}	Urban population (%)	0.2493	0.0046
v_{13}	Area of agricultural land (??)	0.0129	-0.4329

Table 2: Summarization of PCA/SVD loadings

However, higher positive loadings of the variables (v1, v4, v5, v7, and v12) give an idea that it can be one subset of closely related variables with strong positive relationships. Higher negative loadings (v2, v9, v10, and v11) could belong to another subset of closely related variables that have strong negative relationships. Lower positive loading of v13 only and lower negative loadings of (v3, v6 and v8) indicate that neither of the other two subsets of variables is strongly related to them. To summarize, the first PCA vector indicates that there is a strong but negative relationship between “Life Expectancy (years)”, “Adult literacy (% above 15)”, “Mean years of schooling”, “Youth literacy (% age 15-24)” and “Urban population (%)” as one group and “Poverty under $1.25 per day”, “Under 5 mortality”, “Poverty under $2 per day” and “Undernourishment” as another group. In addition, the variables loaded on 1st PCA vector share the same conceptual meaning and are measuring the level of adult and youth literacy, life expectancy and urbanization of observed countries. In other words, 1st PCA component demonstrates social and economical growth of the countries.

Identifying relationships between the variables described by the 2nd principal component vector

Considering the second principal component vector (Table 2 and Fig 1), it is possible to identify a subset with higher negative loadings of variables (v3 and v8) indicating strong negative relationship. We can also observe a subset with lower positive loadings of variables (v5, v6, v9, v11 and v12) along with a subset of lower negative loadings (v1, v2, v4, v7, v10, and v13) but without any significant connections to the other variables. In conclusion, the second PCA component shows roughly equivalent reliance on “Tuberculosis (cases)” and “Population under 15” (as one increases, so does the other). Both variables have negative sign that means that they are not “inversely” correlated. In addition, the variables with significant negative loadings on 2nd PCA vector are measuring the number of tuberculosis cases in population less than 15 years of age.

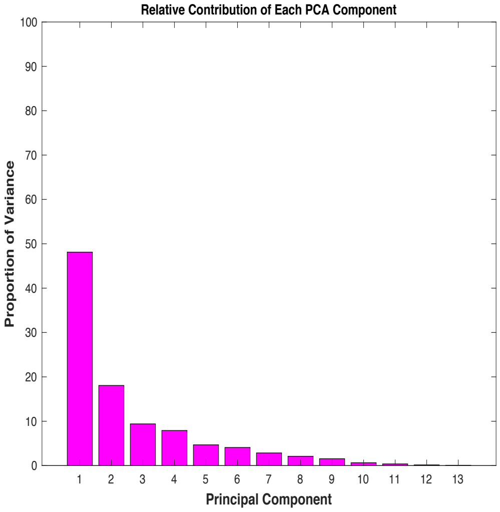
Obtaining singular values and proportions of variation

We aim to find the components that explain the maximum variance because we want to retain as much information as possible using these components. So, higher is the variance (weights), higher will be the information contained in these components. Weights of the PCA are calculated as square roots of sorted singular values from S matrix. To compute the proportion of variance (weights) by each component in percentage, we simply divide the variance by sum of total variance and multiply by 100.

```
% Calculate the weights of the PCA from S matrix
weights = diag(S).^2
% Compute the ratio of each component to the total sum of the weights in percentage
prop_weights = weights / sum(weights) * 100
```

```
prop_weights = 13×1
    48.1127  18.0626  9.3981  7.9103  4.6770  4.0842  2.8522  2.1037  1.5475
```

The scree plot (Fig 2) is used to graphically present the size of the singular values (eigenvalues) related to each component. We can notice that pareto chart (Fig 3) is showing only first seven PCA components that explains 95% of the total variance. In both graphs (Fig 2 and Fig 3), clear breaks in the amount of variance accounted for by each PCA component are between the 1st and the 2nd PCA components, and maybe between 2nd and 3rd components.



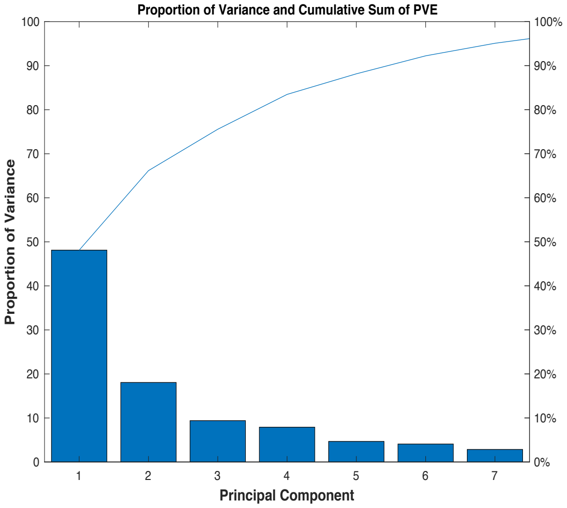


Fig 2 (left): Stacked bar plot demonstrates that the 1st PCA component captures the maximum variance (~48.11%) followed by the rest of the components that capture the remaining variance in the dataset in descending order

Fig 3 (right): Pareto chart displays relative proportion of variances of seven components in descending order by bars with maximum of variance captured by the 1st PCA component, and the cumulative sum of proportion of total variance up to 95% is represented by the line

In our dataset, the singular values that are presented in the matrix S are used to calculate the proportion of variances due to each component which are given as output of *prop_weights* variable above. The 1st PC component alone accounts for ~48.11% of the total variance, the 2nd component accounts for 18.1%, the 3rd component accounts for 9.4%, the 4th component accounts for 7.9%, the 5th component accounts for 4.7%, the 6th component accounts for 4.1% and so on. Each succeeding component accounts for progressively smaller amounts of variance. How many components do we need to retain for interpretation of more than 90% of total variance of the original data set? By analyzing the output (*cumprop_weights* variable below), we can notice that the first 6 principal components explain 92.24% of the total variance, and the rest of the components up to 13 account for only trivial variance. In order words, we can reduce 13 PCA components to 6 PCA components without compromising on proportion of variance. An alternative criterion that can be used to check the obtained results is to retain enough components so the cumulative sum of proportion of variance explained equals to minimal value (Fig 4).

```
% % Calculate the cumulative sum of proprtion of variance (variance explained)
cumprop_weights = cumsum(prop_weights)
```

```
cumprop_weights = 13×1
    48.1127 66.1753 75.5734 83.4837 88.1607 92.2448 95.0970
```

In this study, a critical value for determining the number of components to retain is above 90% (Fig 4). Cumulative sum of proportion of variance explained demonstrates that the cumulative percentage of variance accounted for by components 1, 2, 3, 4, 5 and 6 is 92.24%, as a result, we have to retain all of the first six components. Results of analysis of cumulative sum of proportion of variance explained, as important selective criterion are confirming the findings obtained by very essential criterion assessed previously, both scree plot and percentage of proportion of variance explained.

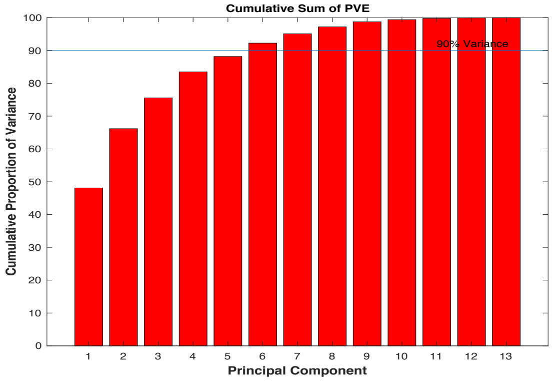


Fig 4: Distribution of cumulative sum of proportion of variance explained (PVE) up to cut off at 90th %

The matrix of scores

```
% Compute the score matrix of first two principal components
T = U*S(:,1:2)
% Plot the 1st and 2nd PCA components of the score matrix
figure
plot(T(:,1),T(:,2), 'ok', 'MarkerEdgeColor', 'k', 'MarkerFaceColor',[.49 1 .63], 'MarkerSize',10)
title('Score Matrix')
xlabel('1st Principal Component', 'FontSize',12, 'FontWeight', 'bold')
ylabel('2nd Principal Component', 'FontSize',12, 'FontWeight', 'bold')
```

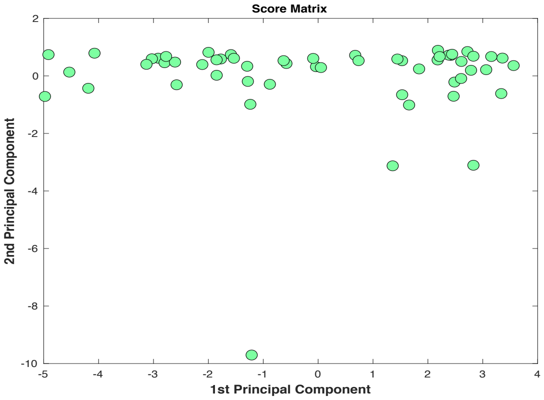


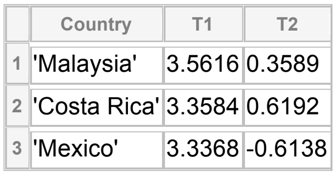
Fig 5: Score matrix displays the mean-centered and scaled data projected onto the first two principal components

```
% Read in the input dataset as a table in order to connect Country variable to PCA components
country_data = readtable("SotTCombined2010.xlsx")
% Perform the same pre-processing steps because dimensions have to be the same
% Remove missing values from the dataset
country_datarm = rmmissing(country_data)
% Extract the first Country column from the modified input data set
country_datarm = country_datarm(:,1)
```

```
% Convert the frist two columns of score matrix into table
T_country = array2table(T)
% Concatenate table T_country and table country_datarm
T_country = [country_datarm T_country]
% Calculating the min values for 1st PCA component
minValue1 = sortrows(T_country,'T1')
% Display of first three countries related to the first three minimal values of 1st component
smallest1 = minValue1([1 2 3],:)
% Calculating the min values for 2nd PCA component
minValue2 = sortrows(T_country,'T2')
% Display of first three countries related to the first three minimal values of 2nd component
smallest2 = minValue2([1 2 3],:)
% Display of first three countries related to the first three maximal values of 1st component
largest1 = minValue1([end end-1 end-2],:)
% Display of first three countries related to the first three maximal values of 2nd component
largest2 = minValue2([end end-1 end-2],:)
```

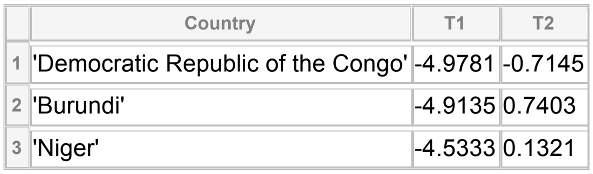
The matrix of scores (T) is the same size as the input data matrix and contains the coordinates of the original data projected in the new coordinate system defined by the principal components. From (Fig 5) by examining magnitude of values in the score matrix shown in the output as minimum and maximum values, Malaysia country has the highest score (3.5616) on the 1st component (Table 3). This is reasonable, considering its dynamic economic and social progress lifted millions of people out of poverty and educational inequality. While on 2nd component, India, Indonesia and Brazil have highest magnitude on y axis of (-9.7036, -3.1258, -3.1061) respectively with negative signs which means that they are not "inversely" correlated (Table 5). These countries can be viewed as extreme countries for 2nd PCA component that interprets India is leading the count, followed by Indonesia in number of people (cases) infected with bacteria *Mycobacterium tuberculosis* and number of children infected under the age of 15. On the other hand, if we observe carefully the score matrix (Fig 5), we can notice that India, Brazil and Indonesia are located at a distance from the amount of the data distribution; as a result, they can be observed as large score matrix outliers (Table 8). If we eliminate them as outliers, then, the next country with the highest magnitude on y-axis is Philippines

Table 3. 1st PCA component: highest magnitude



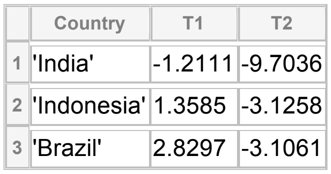
	Country	T1	T2
1	'Malaysia'	3.5616	0.3589
2	'Costa Rica'	3.3584	0.6192
3	'Mexico'	3.3368	-0.6138

Table 4. 1st PCA component: lowest magnitude



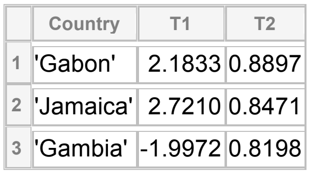
	Country	T1	T2
1	'Democratic Republic of the Congo'	-4.9781	-0.7145
2	'Burundi'	-4.9135	0.7403
3	'Niger'	-4.5333	0.1321

Table 5. 2nd PCA component: highest magnitude



	Country	T1	T2
1	'India'	-1.2111	-9.7036
2	'Indonesia'	1.3585	-3.1258
3	'Brazil'	2.8297	-3.1061

Table 6. 2nd PCA component: lowest magnitude



	Country	T1	T2
1	'Gabon'	2.1833	0.8897
2	'Jamaica'	2.7210	0.8471
3	'Gambia'	-1.9972	0.8198

The minimum values on 1st component is observed at Democratic Republic of Congo (-4.9781), Burundi (-4.9135) and Niger (-4.5333) which means that these countries have the lowest level of Life Expectancy, Adult and Youth Literacy or the lowest economic and social progress (Table 4). Since, the 1st component accounts for the most variation in the dataset, general highest and lowest levels are obtained from 1st PCA component. In general, we can confirm that Malaysia is the most rapid growing country and The Democratic Republic of Congo is the slowest growing country.

A dimensionally reduced representation of the dataset

```
% Compute the reduced matrix from only first six PCA components
reduced = U * S(:,1:6) * V(:,1:6)'
```

```
 reduced = 56×13
    -0.7820     0.5781    -0.2929    -2.0736    -1.4213    -1.1012    -2.0920 ⋯
    -0.5772     0.1274    -0.4048     0.6667     0.9325     1.6614     0.6723
    -1.8073     2.4355    -0.5364    -0.5179    -0.9397    -0.1963    -0.5526
```

The residuals of the reduced representation

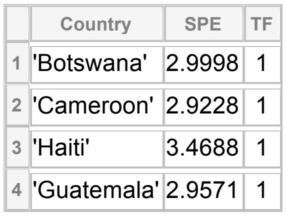
```
% Compute the residuals/SPE. This is the difference between X_scaled and matrix reduced.
% Residuals aid in identifying the outliers in the given data.
SPE = sum((X_scaled - reduced).^2,2)
% Convert SPE into table
spe_country = array2table(SPE)
% Concatenate table spe_country and table country_datarm (previously created above)
spe_country = [country_datarm spe_country]
% Remove outliers with quartiles from SPE
[SPERO,TF] = rmoutliers(SPE,'quartiles')
% Create table for Outliers
Outlier = array2table(TF)
% Concatenate table Outliers with table spe_country
Outlier = [spe_country Outlier]
% Filter Outliers
Outlier(Outlier.TF==1,:)
```

```
SPE = 56×1
     0.3005 2.9998 1.3416 2.9228 0.5057 1.276 1.5974 0.9108 0.3353 0.9161
```

Residuals or SPE (square prediction error) or reconstruction error presents deviation between the reconstructed/reduced data and the original data. It is actually square distance between the original data and the reconstructed data and it is inversely proportional to the total variance of the PCA space. If we select a larger number of components, the variation of lower PCA space will increase and the error between the reconstructed data and the original data will decrease. However, the robustness of PCA is measured by the ratio between the total variance of PCA to the total variance of original data. PCA is very sensitive to pre-processing the data and to outliers. If an observation has a large SPE, then we say this observation is inconsistent within the dataset. PCA components reconstruct the original dataset and provide a correlation structure between the variables in the dataset. When an observation has a large SPE, then that observation is destroying the correlation structure, thus it is inconsistent for the model of PCA. The aim of PCA is dimension reduction that means to explain as much variability (variance) as possible with only a few PCA components. If the data has some underlying data points, the dimension can almost always be reduced. Large outliers increase the non-robustness of PCA or decrease the variation of PCA space. As shown by PCA, there are two types of outliers: large score matrix outliers and residual outliers.

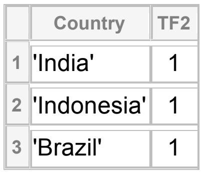
The outlier was determined using the ‘quartile’ method inbuilt in the *rmoutliers ()* function. Four data points obtained as residual outliers are related to Bostwana, Cameroon, Haiti and Guatemala. Score matrix outliers are India, Brazil and Indonesia as shown on Table 7 and Table 8.

Table 7: Residual outliers



	Country	SPE	TF
1	'Botswana'	2.9998	1
2	'Cameroon'	2.9228	1
3	'Haiti'	3.4688	1
4	'Guatemala'	2.9571	1

Table 8: Score matrix outliers on the 2nd PCA component



	Country	TF2
1	'India'	1
2	'Indonesia'	1
3	'Brazil'	1