A Short Monograph on Techniques of Dimension Reduction

TO SERVE AS A REFRESHER FOR PCA

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

INDEX

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

LIST OF FIGURES

LIST OF TABLES

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

1. High Dimensional Data and Dimension Reduction Problem

1.1. What is high dimensional data?

Thanks to modern technology, it is easier than ever to collect and store a huge volume of data. Consider e-commerce like Amazon, a superstore like Big Bazaar, or a car manufacturer like Maruti. If analysis of their customer base is taken on to understand the current requirement, the number of observations in the database might be in millions. Corresponding to each of the observations, there may be several thousand attributes on which data is collected. Each of these data sets is an example of high-dimensional data.

In analytics each attribute is a dimension of the data.

High dimensional data refers to those data sets where the number of attributes is staggeringly high. Even though each dimension, or each attribute, is expected to measure a different aspect of each observation, because all measurements are generated from the same observation, the attributes are highly correlated. Correlated attributes do not contribute to the information in the data set. Further, correlated attributes create instability in the analysis of data. High dimensional data is not informative unless dimensions are orthogonal, *i.e*. uncorrelated or independent.

1.2. Why is high dimension a problem?

Information content, and not necessarily high dimension, of a data set, contribute towards the extraction of better insight from the data. If a variable of interest is associated with several other variables, then the association will make the standard deviation of the estimates of the associated parameters exceptionally high. This may impact the significance of any hypothesis testing. One example of this phenomenon is multicollinearity in regression. Multicollinearity makes the standard deviations of the estimated regression coefficients very high and often reverses the sign of the coefficients. This renders the regression model itself useless.

Unfortunately, pairwise correlation coefficients or scatterplots are not always suitable to identify and eliminate high-dimensionality problems. It is also possible that in the data set more than two attributes are associated. Jointly they influence the analysis but if investigated only pairwise, correlation coefficients may not be high.

Scatterplots typically identify two-dimensional relationships only. It may even be possible to construct and rotate three-dimensional scatterplots to identify whether any three-dimensional association exists in the data. However, it is impossible to meaningfully consider any higher dimensional plots. The sheer number of such pairwise correlation coefficients or scatterplots will be staggering if the total number of attributes is in the hundreds.

Hence it is clear that, before any meaningful analysis of data is undertaken, data dimension must be reduced.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

There are several statistical techniques used to reduce the dimensionality of the data. Two of the most commonly employed methods are the **Principal Component Analysis** and the **Factor Analysis**.

[PCA is covered in DSBA course, but FA is not.]

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

2. Principal Component Analysis (PCA)

2.1. Introduction to Principal Components

The concept of principal components is quite intuitive. Instead of dealing with a large number of possibly correlated variables, principal components are constructed as a suitable linear combination of the observed variables such that the components have two important properties:

- The principal components (PCs) carry the total variance present in the data
- The PCs are orthogonal, *i.e.* uncorrelated, to one another

Information contained in the data is determined by the variance of the attributes. A random variable whose variance is 0, is completely non-informative because for each unit this variable has the same value; in other words, this is a constant. Reduction of dimension involves sacrificing a certain amount of variance. A balance must be struck so that a significant reduction in the number of dimensions is achieved by sacrificing the least possible amount of variance.

2.2. Construction of Principal Components

Before we move forward, let us introduce the notations. Let the observed variables (original attributes) in the data be denoted by $X_1, X_2, \ldots, X_p$ where p is a large number and $Var(X_i) = \sigma_i^2$. Total variance in the data is defined as $\sum_{i=1}^{p} \sigma_i^2$.

Let the principal components be defined by $Y_j, j = 1,2,\ldots,p$. The total number of PCs that can be defined is equal to the number of original attributes in the data. The PCs are linear combinations of the X's and may be defined as

$$Y_1 = w_{11}X_1 + w_{12}X_2 + \ldots + w_{1p}X_p$$

$$Y_2 = w_{21}X_1 + w_{22}X_2 + \ldots + w_{2p}X_p$$

$$\ldots\ldots$$

$$Y_p = w_{p1}X_1 + w_{p2}X_2 + \ldots + w_{pp}X_p$$

where the weights $w_{11}, w_{12}, \ldots, w_{pp}$ need to be determined. In fact the problem of construction of PCs reduces to estimation of $w_{11}, ., w_{pp}$.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Note that the PCs are functions of the random variables $X_1, \ldots, X_p$, and hence they themselves are random variables. Let $Var(Y_j) = \lambda_j, j = 1, \ldots, p$.

The weights w_{ij} are estimated such that

1. $\lambda_1 > \lambda_2 > \ldots > \lambda_p > 0$
2. $\sum_{j=1}^{p} \lambda_j = \sum_{i=1}^{p} \sigma_i^2$
3. Correlation between any pair of PCs is 0.

The property that the first principal component Y_1 has the largest variance, the second principal component Y_2 has the second largest variance and so on, till the p-th PC Y_p has the smallest variance, ensures that PC is a dimension reduction technique.

Total variance of the PCs is equal to the total variance of the original attributes. Since the variances of the PCs are monotonically decreasing, it is possible to use first k PCs, $k < p$, so that a large proportion of total variance is explained by a significantly smaller number of PCs. The number of PCs that are retained is subjective. Typically k is chosen so as to retain 70% - 90% of total variance. The first k principal components $Y_1, Y_2, ., Y_k$ are chosen for further analysis, instead of the original $X_1, \ldots, X_p$, thereby reducing the dimension from p to k, $k < p$.

The smaller k is relative to p, the more reduction in dimension is achieved.

Case Study

The "places" data from the Places Rated Almanac (Boyer and Savageau, 1985) is a collection of 9 composite variables and population constructed for 329 metropolitan areas of the United States. This dataset is taken from the Places Rated Almanac, by Richard Boyer and David Savageau, copyrighted and published by Rand McNally. The composite variables are:

- Climate mildness (climate)
- Housing Cost (housing)
- Health care and environment (healthcare)
- Crime (crime)
- Transportations supply (transport)
- Educational opportunities and effort (education)
- Arts and cultural facilities (arts)
- Recreational opportunities (recreation)
- Personal economics outlook (economics)

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

In addition, 1980 Population for each metropolitan area was also considered.

Each of the 10 attributes associated to the metropolitan areas is measuring different aspects. If a ranking of the metropolitan areas is the objective, then dealing with 10 variables, possibly contradictory, makes the problem very complex. Often a single ranking may not serve any purpose because, to different persons, different aspect of the metropolitan area will be more attractive.

The first goal is to reduce the dimension of the data and investigate what would be an appropriate number of principal components that retain an optimum proportion of variability in the data.

Solution: The objective is to construct the principal components and determine the optimum number of principal components.

2.3. Need for Principal Components

Descriptive Analysis (EDA) on Places Rated Data

```
#Step 1: Import required packages into Jupyter notebook
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import statsmodels.api as sm
from scipy.stats import zscore
from sklearn.decomposition import PCA
from statsmodels.multivariate.pca import PCA
from factor_analyzer import FactorAnalyzer
from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity,calculate_kmo
```

```
#Step 2: Upload data into Jupyter notebook using csv file
places = pd.read_csv ('places_data.csv')

#print the number of rows and columns
places.shape
(329, 11)
```


Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

So, there are in total 329 observations on these 11 values. We now get a glimpse of the first 6 rows of the dataset.

```
#Glimpse of the first 6 rows of the data
places.head(6)
```

	Names	Climate	HousingCost	HlthCare	Crime	Transp	Educ	Arts	Recreat	Econ	Pop
0	Abilene,TX	521	6200	237	923	4031	2757	996	1405	7633	110932
1	Akron,OH	575	8138	1656	886	4883	2438	5564	2632	4350	660328
2	Albany,GA	468	7339	618	970	2531	2560	237	859	5250	112402
3	Albany-Schenectady-Troy,NY	476	7908	1431	610	6883	3399	4655	1617	5864	835880
4	Albuquerque,NM	659	8393	1853	1483	6558	3026	4496	2612	5727	419700
5	Alexandria,LA	520	5819	640	727	2444	2972	334	1018	5254	135282

Except for Population, other values are composite scores.

Since no numerical analysis can be performed on the first column, it is stored in a separate variable called “names” and is being deleted from “places” data for ease of analysis. All the other variables in the data are continuous.

PCA should be performed with continuous variables only.

```
#Collect the names of metro areas separately and view the first 6 names
names=places['Names']
names.head(6)
0                Abilene,TX
1                 Akron,OH
2                 Albany,GA
3    Albany-Schenectady-Troy,NY
4              Albuquerque,NM
5               Alexandria,LA
Name: Names, dtype: object

# Delete the first column containing names from the dataset
places = places.drop('Names',axis = 1)
```

Basic descriptive analysis of all variables is performed.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

#Find the arithmetic average of the variables

pd.DataFrame(round(places.mean(),2)).T

Climate	HousingCost	HlthCare	Crime	Transp	Educ	Arts	Recreat	Econ	Pop
538.73	8346.56	1185.74	961.05	4210.08	2814.89	3150.88	1845.96	5525.36	522118.45

Variance-covariance matrix of the variables is shown below. The variances are on the main diagonal.

#Find variance covariance matrix

round(places.cov(),2)

	Climate	HousingCost	HlthCare	Crime	Transp	Educ	Arts	Recreat	Econ	Pop
Climate	14594.64	111313.31	25846.07	8300.97	13870.87	2500.43	127293.08	20838.39	-13112.12	26407923.21
HousingCost	111313.31	5689477.78	1083790.89	114344.31	941240.94	151454.13	4967020.42	813760.45	696953.13	695269315.24
HlthCare	25846.07	1083790.89	1006013.08	109137.04	684563.30	157735.73	4031336.88	263673.54	75347.45	741496988.98
Crime	8300.97	114344.31	109137.04	127559.11	148532.09	8526.06	645766.48	99438.76	100701.79	131683250.83
Transp	13870.87	941240.94	684563.30	148532.09	2105921.19	156413.94	3131295.53	427589.88	93240.06	521419005.44
Educ	2500.43	151454.13	157735.73	8526.06	156413.94	102908.12	555891.51	20164.89	41642.54	102945093.46
Arts	127293.08	4967020.42	4031336.88	645766.48	3131295.53	555891.51	21550798.30	1420141.86	380970.47	3684692829.27
Recreat	20838.39	813760.45	263673.54	99438.76	427589.88	20164.89	1420141.86	652683.30	152035.16	266958609.86
Econ	-13112.12	696953.13	75347.45	100701.79	93240.06	41642.54	380970.47	152035.16	1176071.98	57892734.26
Pop	26407923.21	695269315.24	741496988.98	131683250.83	521419005.44	102945093.46	3684692829.27	266958609.86	57892734.26	798025356821.36

Let us focus on the variances. Whereas $\hat{\sigma}_1^2 = 14594.64$ (climate), $\hat{\sigma}_5^2 = 2105921.19$ (transport) and $\hat{\sigma}_{10}^{\ 2} = 798025356821.36$. When variances are so widely different, it is not a good idea to perform PCA on the unscaled variables. PCA works on the total variance which is the sum of the variances in the data. If one variance (or more) variance(s) is (are) very high compared to the rest, it (they) will dominate the construction of the PCs and all variables will not have proper representation.

Note that it is important to check the sample size used by Python as the divisor while calculating variance and covariance. Using the wrong divisor will provide a biased estimate of the population variance. Setting the parameter *ddof = 1* ensures that the divisor is $n - 1$.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

When sample variances of the original variables show differences by large order of magnitude, variables need to be normalized.

Define $Z_i = \frac{X_i - \overline{X}_i}{sd(X_i)}$, for $i = 1,2,\ldots,p$. Then each Z_i has mean 0 and variance 1. Total variance in the data is p.

PCA is performed on the scaled variables, instead of on the original variables.

```
# Normalize the variables and then view the first 6 rows
std_places = pd.DataFrame(zscore(places,ddof=1),columns=places.columns)
np.round(std_places.head(6),2)
```

	Climate	HousingCost	HlthCare	Crime	Transp	Educ	Arts	Recreat	Econ	Pop
0	-0.15	-0.90	-0.95	-0.11	-0.12	-0.18	-0.46	-0.55	1.94	-0.46
1	0.30	-0.09	0.47	-0.21	0.46	-1.17	0.52	0.97	-1.08	0.15
2	-0.59	-0.42	-0.57	0.03	-1.16	-0.79	-0.63	-1.22	-0.25	-0.46
3	-0.52	-0.18	0.24	-0.98	1.84	1.82	0.32	-0.28	0.31	0.35
4	1.00	0.02	0.67	1.46	1.62	0.66	0.29	0.95	0.19	-0.11
5	-0.16	-1.06	-0.54	-0.66	-1.22	0.49	-0.61	-1.02	-0.25	-0.43

Note that *np.round()* function has been used to display the numerical values upto 2 places of decimal only.

Scaling ensures that attribute means are all 0 and variances 1.

```
# Find new variance-covariance matrix of the transformed variables
np.round(std_places.cov(),2)
```


Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

	Climate	HousingCost	HlthCare	Crime	Transp	Educ	Arts	Recreat	Econ	Pop
Climate	1.00	0.39	0.21	0.19	0.08	0.06	0.23	0.21	-0.10	0.25
HousingCost	0.39	1.00	0.45	0.13	0.27	0.20	0.45	0.42	0.27	0.33
HlthCare	0.21	0.45	1.00	0.31	0.47	0.49	0.87	0.33	0.07	0.83
Crime	0.19	0.13	0.31	1.00	0.29	0.07	0.39	0.35	0.26	0.41
Transp	0.08	0.27	0.47	0.29	1.00	0.34	0.47	0.37	0.06	0.40
Educ	0.06	0.20	0.49	0.07	0.34	1.00	0.37	0.08	0.12	0.36
Arts	0.23	0.45	0.87	0.39	0.47	0.37	1.00	0.38	0.08	0.89
Recreat	0.21	0.42	0.33	0.35	0.37	0.08	0.38	1.00	0.17	0.37
Econ	-0.10	0.27	0.07	0.26	0.06	0.12	0.08	0.17	1.00	0.06
Pop	0.25	0.33	0.83	0.41	0.40	0.36	0.89	0.37	0.06	1.00

Note that all variances are now 1 (main diagonal). In fact, this matrix is same as the correlation matrix of the original (unscaled) variables. [You may verify that the output of ***np.round(std_places.corr(),2)*** is identical]

We will work with the "std_places" data frame from now on. To investigate association between the variables visually, scatterplots of all possible pairs of the variables are considered.

```
#Scatterplots of all possible variable pairs
def hide_current_axis(*args, **kwds):
    plt.gca().set_visible(False)
g = sns.pairplot(std_places)
g.map_diag(hide_current_axis)
```


Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

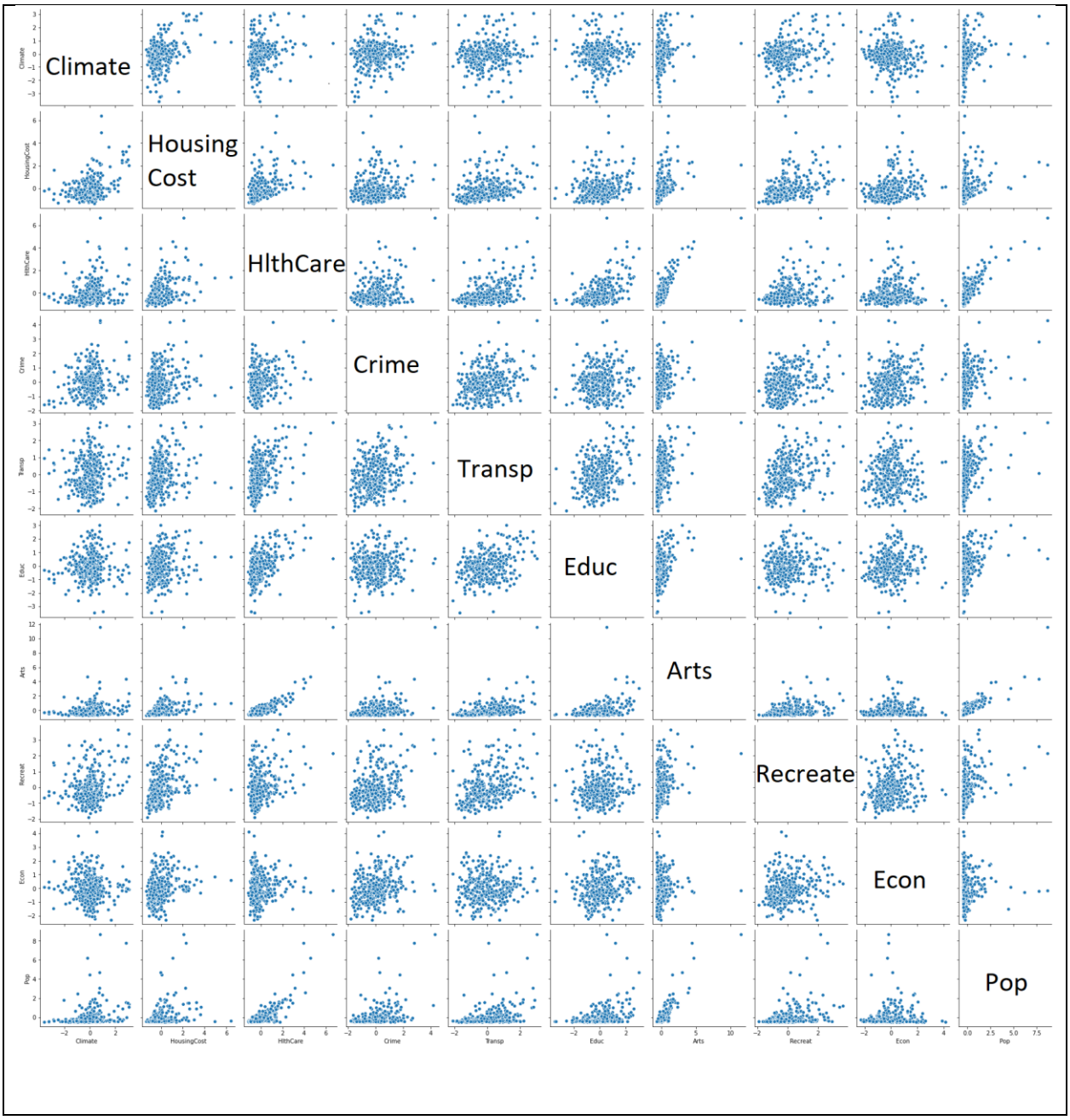


Fig. 1. Pairwise plots of the variables

The correlation matrix and pairwise scatterplots indicate high correlation among *population*, *arts* and *healthcare*. Moderate correlation may be detected between several pairs of variables, such as *housing cost*, *arts*, *recreation* and *healthcare*; between *transport* and *healthcare*; etc. Existence of such pairs of high and moderate correlations indicate that dimension reduction must be considered for the Places Rated data.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

2.4. Optimum Number of Principal Components

All principal components are extracted at one go and then optimum number of components decided.

```
# Principal Component Extraction using sklearn.decomposition package
pca = PCA(n_components= 10)
pca.fit_transform(std_places)

pc_comps = ['PC1','PC2','PC3','PC4','PC5','PC6','PC7','PC8','PC9','PC10']
prop_var = np.round(pca.explained_variance_ratio_,2)
std_dev = np.round(np.sqrt(pca.explained_variance_),2)
cum_var = np.round(np.cumsum(pca.explained_variance_ratio_),2)
temp = pd.DataFrame(pc_comps,columns=['PCs'])
temp['Proportion Of Variance'] = prop_var
temp['Standard Deviation'] = std_dev
temp['Cumulative Proportion'] = cum_var
temp
```

PCs	Proportion Of Variance	Standard Deviation	Cumulative Proportion
PC1	0.41	2.02	0.41
PC2	0.13	1.12	0.54
PC3	0.11	1.07	0.65
PC4	0.10	0.98	0.74
PC5	0.08	0.89	0.82
PC6	0.07	0.83	0.89
PC7	0.05	0.70	0.94
PC8	0.03	0.59	0.98
PC9	0.01	0.37	0.99
PC10	0.01	0.31	1.00

Recall that there are 10 observed variables $X_1, \ldots, X_{10}$, (or their scaled version $Z_1, \ldots, Z_{10}$) and hence, 10 PCs are generated. The principal components are constructed in decreasing order of magnitude of their standard deviations, which is equivalent to decreasing order of magnitude of

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

their variances. Total variance of the scaled variables is 10. In Table 1 the variances of the constructed principal components and their sum total is given.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	SUM
SD	2.02	1.12	1.07	0.98	0.89	0.83	0.70	0.59	0.37	0.31	
Var (λ)	4.09	1.26	1.14	0.95	0.79	0.69	0.50	0.34	0.14	0.09	10.00

Table 1: Standard deviation and variances of the principal components

The proportion of variance of a principal component is obtained by dividing the variance of the component (obtained by squaring the standard deviation), by total variance. The cumulative proportion up to the *k*-th principal component is the sum of the proportions of variances up to the k-th component, i.e. $\sum_{j=1}^{k} \lambda_j$.

If $k = 5$, cumulative proportion is 82.37%. Although there are 10 observed variables, the first 5 principal components can explain more than 80% of the total variation. Hence it is sufficient to use the first 5 PCs instead of the original 10 variables, thereby reducing the dimensions by half.

The optimum choice of k is subjective. The set of k principal components effectively substitute the original p variables. General rule of thumb is to choose k so as to explain 70% - 90% of the total variance. Often a screeplot is used to determine k.

```
# Obtain the screeplot
plt.figure(figsize=(10,5))
plt.plot(temp['Proportion Of Variance'],marker = 'o')
plt.xticks(np.arange(0,11),labels=np.arange(1,12))
plt.xlabel('# of principal components')
plt.ylabel('Proportion of variance explained')
```


Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

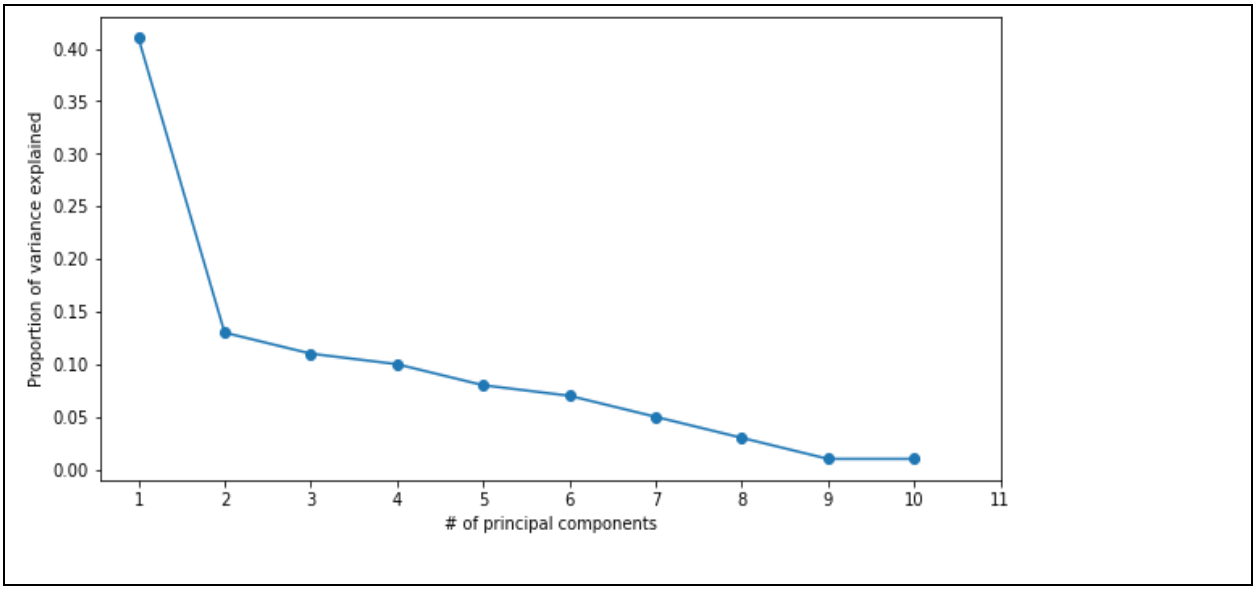


Fig. 2. Scree plot done on PCA

The scree plot is a useful visual tool to select k. On the X-axis are shown the indices of the PCs and on the Y-axis are shown the variances. If there is a distinct break point in the line joining the variances (elbow point) beyond which the line becomes approximately horizontal, then that point may be taken as the value of k, provided other conditions are also satisfied.

In Fig 2, there is a distinct break at 2. However, k cannot be taken to be 2 since the first two PCs explain only 53% of total variance. The PCs must be taken so as to explain between 70% - 90% of the total variance. If $k = 4$, then the first 4 PCs explain 74% of the total variance. One choice of k could have been 4. However, we have taken $k = 5$ so that the explained variance is above 80%.

2.5. Structure of Principal Components and PC Scores

Principal components are linear combinations of the original variables. Each PC is a linear combination of all variables, or scaled variables, as the case may be. It is possible that some of the coefficients are very small numbers or close to 0. We present the linear combinations that make up the first 5 PC's.

```
# Print first 5 PCs
pc_df_pcafunc =
pd.DataFrame(np.round(pca.components_,2),index=pc_comps,columns=std_places.columns)
pc_df_pcafunc.head(5)
```


Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

	Climate	HousingCost	HlthCare	Crime	Transp	Educ	Arts	Recreat	Econ	Pop
PC1	0.18	0.30	0.44	0.25	0.30	0.25	0.45	0.28	0.10	0.43
PC2	-0.21	-0.34	0.26	-0.33	0.08	0.36	0.18	-0.42	-0.53	0.20
PC3	0.70	0.21	-0.01	-0.16	-0.17	-0.27	0.02	0.07	-0.58	0.02
PC4	0.14	0.52	0.05	-0.58	-0.09	0.46	-0.10	-0.13	0.29	-0.20
PC5	0.22	-0.07	0.13	0.26	-0.67	0.01	0.16	-0.50	0.30	0.24

For each PC, the row of length 10 gives the weights with which the corresponding variables need to be multiplied to get the PC. Note that the weights can be positive or negative. So, for example,

PC1 = 0.18 * SClimate + 0.3 * SHousingCost + 0.44 * SHlthCare + 0.25 * SCrime + 0.3 * STransp + 0.25 * SEduc + 0.45 * SArts + 0.28 * SRecreat + 0.1 * SEcon + 0.43 * SPop

PC2 = -0.21 * SClimate - 0.34 * SHousingCost + 0.26 * SHlthCare - 0.33 * SCrime + 0.08 * STransp + 0.36 * SEduc + 0.18 * SArts - 0.42 * SRecreat - 0.53 * SEcon + 0.20 * SPop

Table 2: PC1 and PC2 expressed in terms of scaled variables

The letter S indicates that the scaled (normalized) variable is used to construct the PCs.

Similarly, the other PCs can also be expressed in terms of the scaled variables.

Once the original variables are replaced by the PCs, the latter are used for any further analysis. Just as each observed unit has a particular value of each variable, similarly each observation has a particular value for each PC. These values are called PC scores.

These scores are obtained by putting scaled values of the variables in the expression of PCs as shown in Table 2. [*Hand calculation of PC scores are provided separately*]

```
# Find PC scores
pc = pca.fit_transform(std_places)
pca_df = pd.DataFrame(pc,columns=pc_comps)
np.round(pca_df.iloc[:6,:5],2)
```


Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

	PC1	PC2	PC3	PC4	PC5
0	-1.18	-0.92	-1.38	0.23	0.64
1	0.49	0.06	1.17	-0.97	-0.94
2	-1.86	0.17	-0.04	-0.36	0.93
3	0.97	1.45	-1.24	1.12	-1.17
4	1.87	-0.63	-0.03	-0.59	-0.79
5	-1.77	0.91	-0.10	0.32	0.85

To check that the PCs are orthogonal, correlation matrix is computed.

#Correlation matrix of PC scores

round(pca_df.corr(),2)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
PC1	1.0	-0.0	0.0	0.0	0.0	-0.0	-0.0	0.0	-0.0	0.0
PC2	-0.0	1.0	0.0	0.0	-0.0	-0.0	-0.0	0.0	-0.0	-0.0
PC3	0.0	0.0	1.0	0.0	-0.0	0.0	0.0	0.0	0.0	-0.0
PC4	0.0	0.0	0.0	1.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.0
PC5	0.0	-0.0	-0.0	0.0	1.0	-0.0	-0.0	-0.0	0.0	0.0
PC6	-0.0	-0.0	0.0	-0.0	-0.0	1.0	0.0	-0.0	0.0	-0.0
PC7	-0.0	-0.0	0.0	-0.0	-0.0	0.0	1.0	0.0	-0.0	-0.0
PC8	0.0	0.0	0.0	-0.0	-0.0	-0.0	0.0	1.0	-0.0	-0.0
PC9	-0.0	-0.0	0.0	-0.0	0.0	0.0	-0.0	-0.0	1.0	0.0
PC10	0.0	-0.0	-0.0	-0.0	0.0	-0.0	-0.0	-0.0	0.0	1.0

Let us now investigate the correlations among the first 5 PCs with the original 10 variables.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

result = pd.concat((std_places,pca_df),axis = 1).corr()
np.round(result.iloc[0:10,10:15],2)

	PC1	PC2	PC3	PC4	PC5
Climate	0.36	-0.24	0.75	0.13	0.19
HousingCost	0.60	-0.38	0.22	0.51	-0.06
HlthCare	0.88	0.29	-0.01	0.05	0.11
Crime	0.51	-0.37	-0.17	-0.57	0.23
Transp	0.62	0.09	-0.18	-0.09	-0.59
Educ	0.50	0.40	-0.29	0.45	0.01
Arts	0.90	0.20	0.02	-0.10	0.14
Recreat	0.57	-0.47	0.07	-0.13	-0.44
Econ	0.21	-0.59	-0.62	0.28	0.27
Pop	0.87	0.23	0.02	-0.19	0.21

Though principal components do not necessarily carry any intuitive interpretation, often it is easier to understand if they do. Among the correlations between the PCs and the constituent variables, the following are considerably large:

- PC1 and (health care, arts, population)
- PC3 and climate

Note that while considering the correlations, only numerical values are considered. In the next section these associations will be exploited to investigate whether any underlying *factors* can be defined.

2.6. Further considerations on Principal Components

2.6.1. More on optimum number of principal components

Choosing the correct number of principal components is pivotal in data analysis and requires a balancing act. On one side, the aim is to reduce the dimension, so keeping too many principal

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

components will not serve the purpose. However, keeping too few components will cause a large proportion of total variation among the original variables to remain unexplained.

It has already been noted (Sec 2.4) that choosing the optimum number of principal components is subjective and there is no universal answer to this question. As a general rule of thumb, we keep the first k principal components (out of p, the total number of variables and hence the total number of principal components) that together explain about 70% - 90% of the variation amongst the data.

If the first k PC's satisfy this property, and also if going from k-th to $k + 1$-th PC the cumulative proportion of variance increases marginally (say less than 10%), then the first k PCs are considered but not the $k + 1^{st}$.

Alternatively, If the first k PC's explain a cumulative proportion of variance just above 70%, and the inclusion of the $k + 1^{st}$.PC explains about 80% of the variation albeit the increase in cumulative proportion of variance due to inclusion of $k + 1^{st}$ is less than 10%, we would prefer to keep the first $k + 1$ PC's.

Another rule of thumb (Kaiser Criterion) is not to include any PC if its variance is less than 1. Recall that all scaled variables (Z's) have variance equal to 1. The first few principal components are expected to have higher variances. In Places Rated data, PC1 – PC3 all have variances greater than 1.

2.6.2. Can PCA be used when observed variables are not continuous?

If the variables are not continuous but categorical, it is strongly advised **not** to use PCA. For categorical variables, the usual notions of mean and variance do not work since the values taken by the variables are simply labeled on which the usual rules of addition, subtraction, multiplication, and division do not apply. We have seen that PCA is dependent upon variance which has absolutely no meaning when it comes to categorical variables.

However, there are two kinds of categorical variables: Nominal and Ordinal.

Nominal variables are those for which we do not have any notion of ordering among the values taken by the variable.

For nominal variables, PCA should never be used.

Ordinal variables are those for which we have a notion of ordering among the values. Usually we give the labels to ordinal variables respecting the hierarchy. For example, *socio-economic status* may be defined to be an ordinal variable with three labels: 0 for low, 1 for medium, and 2 for high.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Note that we have an inherent order among the values: $0 < 1 < 2$ reflecting the ordering among the socio-economic levels.

For ordinal variables, however, PCA has been used in the literature, particularly in socio-economic studies. We mention here two references:

(1) "*Estimating Wealth Effects without Expenditure Data – or Tears: An Application to Educational Enrollments in States of India*" by Deon Filmer and Lant Pritchett (1998)

(2) "*The Use of Discrete Data in Principal Component Analysis for Socio-Economic Status Evaluation*" by Stanislav Kolesnikov and Gustavo Angeles (2005)

Although ideally use of PCA should be restricted to continuous variables, an ordinal variable may be considered a surrogate for a hidden continuous variable. The hierarchical values of an ordinal variable represent discrete measurements on the underlying continuous variable.

For example, socio-economic status is often decided by the earnings of an individual (measurable) and other non-measurable characteristic. Earnings is a continuous variable. If the earnings exceed a certain threshold, the person is said to have "high" socio-economic status. If the earnings falls below another threshold, the person has "low" socio-economic status. If the earnings lies between these two thresholds, then the person is said to have "medium" socio-economic status. Although the variable socio-economic status itself is ordinal, it serves as a proxy for the underlying hidden variable earnings, and therefore performing PCA with it is somewhat justified.

Justification must be given in a case-by-case basis when all variables are not continuous.

2.6.3. Is PCA always expected to reduce dimension?

In the Places Rated data, although there were 10 original attributes, more than 80% of the total variance can be explained with only the first 5 PC's, and thus the goal of dimension reduction was achieved.

However, it is never guaranteed that PCA will reduce dimensions for all data. It works only when a large number of original variables are highly correlated with each other. In real data, while dealing with hundreds of variables, usually one finds that many of the variables are associated with each other. PCA works well in such situations.

If on the other hand, there is very weak dependence among the variables, PCA will not help in dimension reduction.

2.6.4. Alternative code to extract PCA

In Python alternative procedure exists to extract principal components.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

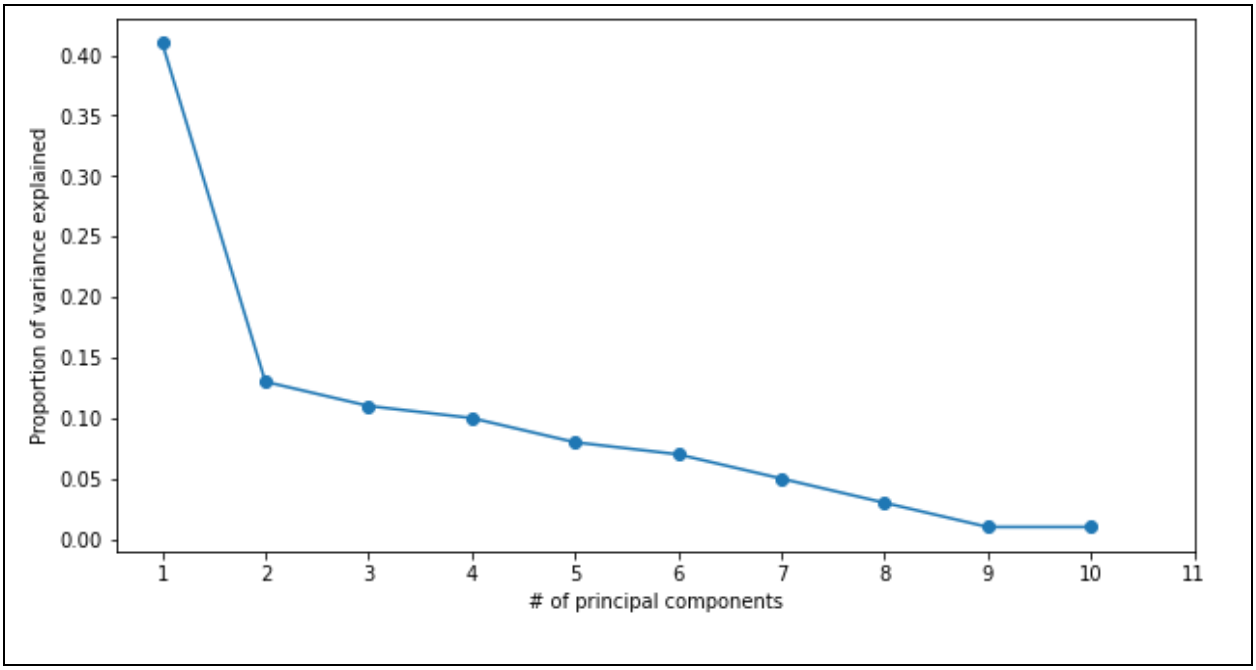
Package `sklearn.decomposition` has been used to obtain the PCs in the current case. There is an alternative package statsmodels.multivariate.pca that performs similarly.

```
# Extraction of PCA with statsmodels.multivariate.pca package
pc = PCA(std_places,method='eig')
cum_var = np.round(pc.rsquare[1:],2)
cum_var.reset_index(drop=True,inplace = True)
var_exp = np.round(pc.eigenvals/np.sum(pc.eigenvals),2)
measure_df = pd.DataFrame(pc_comps,columns=['PCs'])
measure_df['Cumulative Proportion'] = cum_var
measure_df['Proportion of Variance'] = var_exp
measure_df
```

	PCs	Cumulative Proportion	Proportion of Variance
0	PC1	0.41	0.41
1	PC2	0.54	0.13
2	PC3	0.65	0.11
3	PC4	0.74	0.10
4	PC5	0.82	0.08
5	PC6	0.89	0.07
6	PC7	0.94	0.05
7	PC8	0.98	0.03
8	PC9	0.99	0.01
9	PC10	1.00	0.01

```
# Obtain the screeplot
plt.figure(figsize=(10,5))
plt.plot(measure_df['Proportion of Variance'],marker = 'o')
plt.xticks(np.arange(0,11),labels=np.arange(1,12))
plt.xlabel('# of principal components')
plt.ylabel('Proportion of variance explained')
```


Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.



Compare the above with output of sklearn.decomposition in Sec 2.4 and check that both outputs are very similar though not identical. We take as before, the first 5 PCs. Next, we view the linear combinations corresponding to the first 5 PCs.

```
# Coefficients of PCA
loadings = pd.DataFrame(np.round(pc.loadings,2))
loadings.iloc[:,:5]
```


Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

	comp_0	comp_1	comp_2	comp_3	comp_4
Climate	-0.18	0.21	0.70	-0.14	-0.22
HousingCost	-0.30	0.34	0.21	-0.52	0.07
HlthCare	-0.44	-0.26	-0.01	-0.05	-0.13
Crime	-0.25	0.33	-0.16	0.58	-0.26
Transp	-0.30	-0.08	-0.17	0.09	0.67
Educ	-0.25	-0.36	-0.27	-0.46	-0.01
Arts	-0.45	-0.18	0.02	0.10	-0.16
Recreat	-0.28	0.42	0.07	0.13	0.50
Econ	-0.10	0.53	-0.58	-0.29	-0.30
Pop	-0.43	-0.20	0.02	0.20	-0.24

We find that the linear combinations are identical (up to many places of decimal).

The principal component analysis is usually used as an intermediary step to further analysis. The outcome of PCA may be used in regression, such as principal component regression or partial least squares regression. It can also be used in the extraction of latent factors, which often provide important insights into various business applications, such as customer behavior, e-commerce etc.

Appendix

1. Places Rated PCA Score.xlsx: Contains data and calculation of PCA Scores

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

References:

Anderson, T. W. (2003). An Introduction to Multivariate Statistical Analysis. Wiley Series in Probability and Statistics.

Brian, E. & Torsten, H. (2011). An Introduction to Applied Multivariate Analysis with R. Springer

Johnson, R. A. & Wichern, D. W. (2002). Applied Multivariate Statistical Analysis. Prentice Hall, New Jersey.

Mardia, K. V., Kent, J. T. & Bibby, J. M. (1979). Multiariate Analysis. Academic Press Ltd., London.

Rencher, A. C. (2002). Methods of Multivariate Analysis. Wiley Series in Probability and Statistics.

Becker, R. A., Denby, L., McGill, R. & Wilks A. R. (1987). Analysis of Data from the Places Rated Almanac. The American Statistician,Vol. 41, No. 3 (Aug., 1987), pp. 169-186.

Filmer, D. & Pritchett, L. (1998). Estimating Wealth Effects without Expenditure Data – or Tears: An Application to Educational Enrollments in States of India. Demography, Vol. 38, No. 1 (Feb., 2001), pp. 115-132.

Kolenikov, S. & Angeles, G. (2005). The Use of Discrete Data in Principal Component Analysis for Socio-Economic Status Evaluation. Chapel Hill: Carolina Population Center, University of North Carolina, 1-59.

https://newonlinecourses.science.psu.edu/stat505/

https://nptel.ac.in/courses/111104024/

https://www.statistics.com/multivariate-statistics/

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.