greatlearning
Learning for Life

A Short Monograph on Techniques of Dimension Reduction

PRINCIPAL COMPONENT ANALYSIS
FACTOR ANALYSIS

TO SERVE AS A REFRESHER FOR PGP-DSBA

INDEX

LIST OF FIGURES

LIST OF TABLES

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 huge volume of data. Consider an ecommerce 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 data base 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, due to the fact that all measurements are generated from the same observation, the attributes are highly correlated. Correlated attributes do not contribute to 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 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 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 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.

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. Nevertheless, it is included in the Monograph to make it comprehensive]

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 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 content 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 certain amount of variance. A balance must be struck so that 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 $w_{11}, ., w_{pp}$.

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)

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 retains 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 import multivariate
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)
```

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 delete 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.

```
#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

np.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 Python uses the sample size *n-1* as the default divisor while calculating the covariance.

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)
```

	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)
```

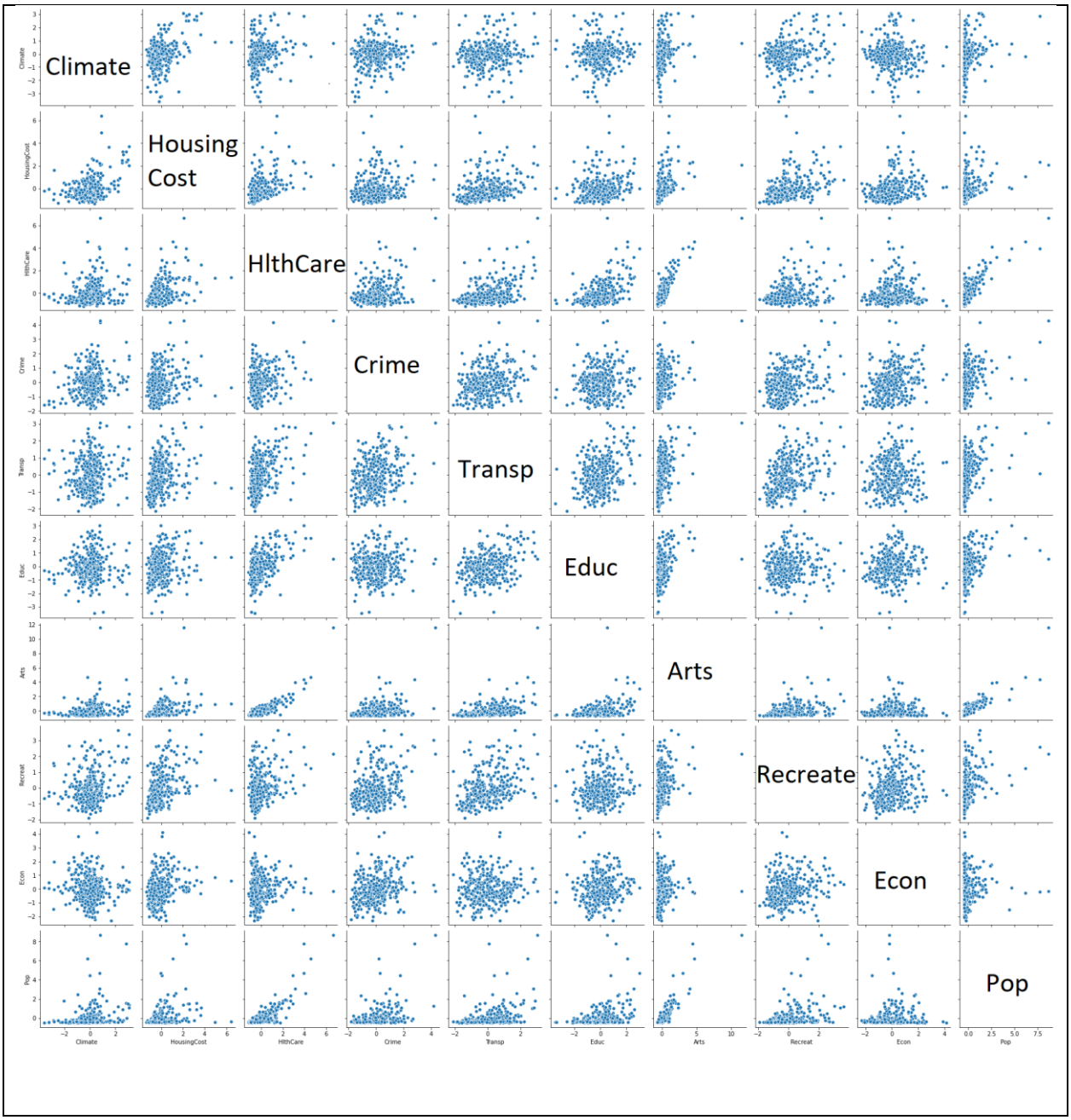


Fig. 1. Pairwise plots of the variables

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.

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 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 upto the k-th principal component is the sum of the proportions of variances upto 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')
```

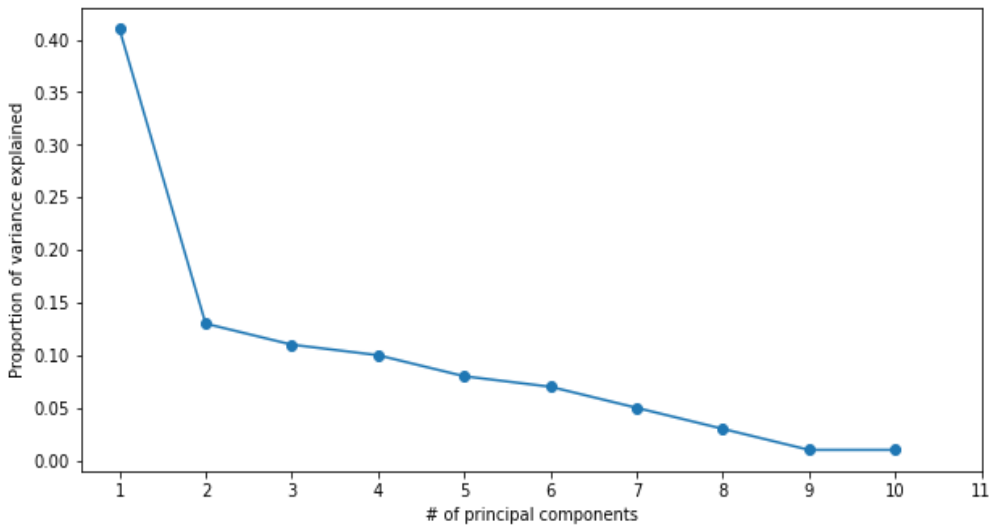


Fig. 2. Scree plot done on PCA

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)
```

	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)
```

	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	1.00	-0.00	0.00	0.00	-0.00	0.00	-0.00	-0.00	-0.00	0.00
PC2	-0.00	1.00	-0.00	-0.00	0.00	-0.00	-0.00	0.00	-0.00	-0.00
PC3	0.00	-0.00	1.00	0.00	0.00	-0.00	-0.00	0.00	-0.00	-0.00
PC4	0.00	-0.00	0.00	1.00	-0.00	-0.00	0.00	-0.00	-0.00	0.00
PC5	-0.00	0.00	0.00	-0.00	1.00	-0.00	0.00	0.00	-0.00	0.00
PC6	0.00	-0.00	-0.00	-0.00	-0.00	1.00	-0.00	0.00	-0.00	-0.00
PC7	-0.00	-0.00	-0.00	0.00	0.00	-0.00	1.00	-0.00	-0.00	0.00
PC8	-0.00	0.00	0.00	-0.00	0.00	0.00	-0.00	1.00	0.00	-0.00
PC9	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	0.00	1.00	-0.00
PC10	0.00	-0.00	-0.00	0.00	0.00	-0.00	0.00	-0.00	-0.00	1.00

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

```
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 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 labels on which the usual rules of addition, subtraction, multiplication and division do not apply. We have seen that PCA is dependent upon variance which have 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. 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

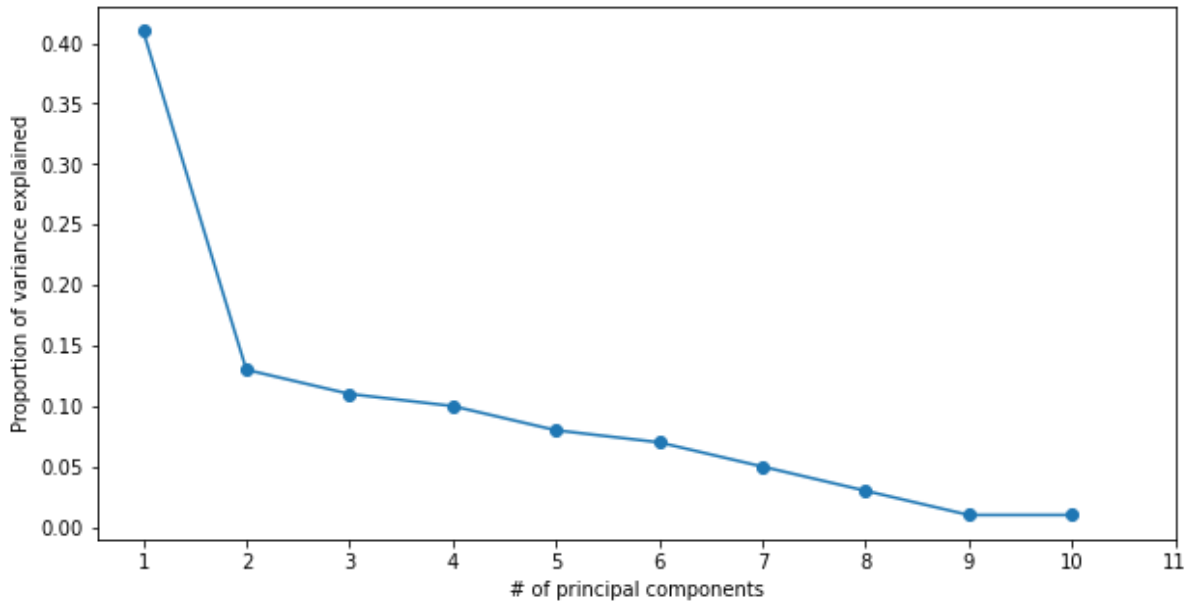
In Python alternative procedure exists to extract principal components.

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 = multivariate.pca.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')
```



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]
```

	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).

Principal component analysis is usually used as an intermediary step to further analysis. Outcome of PCA may be used in regression, such as principal component regression or partial least squares regression. It can also be used in extraction of latent factors, which often provides important insights into various business applications, such as customer behavior, ecommerce etc. This is known as Factor Analysis, which is the topic of our next section.

3. Factor Analysis (FA) - [Self-Exploratory Topic]

Factor analysis is another dimension reduction technique, where, instead of linearly combining $X_1, X_2, \ldots, X_p$, into $Y_1, Y_2, \ldots, Y_k$, a set of underlying factors are identified.

An example will simplify the concepts. Suppose an IQ Test is being administered to a group of young adults entering a college. Among many different types of intelligence, it has been decided to measure inductive, deductive, verbal, numerical and memory. In addition, there is a general or common intelligence factor.

Note that none of these intelligence factors are directly measurable. Hence these are the latent factors. There are many questions given in any IQ test, which are devised to measure these various types of IQ. Scores on these questions are the observed attributes$X_1, X_2, \ldots, X_p$. Identification and scores on the *m* latent factors (*m* = 5, in this case) is the objective of factor analysis.

The underlying key concept of factor analysis is that multiple observed variables have similar patterns of responses because they are all associated with a latent (i.e. not directly measured) variable or factor. Factors typically are broad concepts or ideas that may describe an observed phenomenon.

Like PCA, objective of FA is to preserve as much of the total variance in the data as possible through the latent factors. Theoretically speaking, it is possible to define as many latent factors as the number of variables. However, that way neither the purpose of dimension reduction, not identification of latent variables, is served. Hence we will not consider that situation.

3.1. Identification of Latent Factors

Factor analysis is a method for modeling observed variables, and their covariance structure, in terms of a smaller number of underlying unobservable (latent) factors.

Observed variables can be written as a linear combination of latent factors and an error term

$$X_1 - \mu_1 = \gamma_{11}f_1 + \gamma_{12}f_2 + \ldots + \gamma_{1m}f_m + \varepsilon_1$$

$$X_2 - \mu_2 = \gamma_{21}f_1 + \gamma_{22}f_2 + \ldots + \gamma_{2m}f_m + \varepsilon_2$$

$$X_3 - \mu_3 = \gamma_{31}f_1 + \gamma_{32}f_2 + \ldots + \gamma_{3m}f_m + \varepsilon_3$$

$$\ldots \quad \ldots \quad \ldots$$

$$X_p - \mu_p = \gamma_{p1}f_1 + \gamma_{p2}f_2 + \ldots + \gamma_{pm}f_m + \varepsilon_p$$

As before, $X_1, X_2, \ldots, X_p$, denote the observed variables and let $\mu_1, \mu_2, \ldots, \mu_p$ denote their means respectively. The observed variables can be written as linear combinations of *m* common factors f_1, f_1, …, f_m and *p* additional unique factors or errors. The common factors are uncorrelated random variables; ε_i (i=1, 2, …, p) are error variables, or unique factors, uncorrelated with each other and with f_j and represent the residual error due to the use of common factors.

Total variance in the data can be partitioned into variability due to the common factors and error variance. Let us assume Var(f_j) = 1, for all $j = 1, \ldots, m$. Hence

$$Var(X_i) = \sigma_{ii} = \sum_{j=1}^{m} \gamma_{ij}^2 + \Psi_i^2$$

where $h_i^2 = \sum_{j=1}^{m} \gamma_{ij}^2$ is the **Communality** of *i*-th variable. Communality is the proportion of variance of X_i explained by all the latent factors taken together. Ψ_i^2 = Var(ε_i) is called the **uniqueness** of the factor. Naturally, factor analysis is meaningful when communality is higher than uniqueness. If a variable does not relate to any of the latent factors, and stand apart, its communality will be very small and uniqueness very high.

3.2. Main Differences between PCA and FA

Note that the main difference between PCA and FA are

(1) PCA is an algorithm and FA is one of its many applications. Other applications of PCA include principal component regression, partial least square regression, reduction of multicollinearity etc.

(2) In PCA the observed variables are combined to reduce dimension but in FA the underlying latent variables are identified and extracted. Factors are expected to be intuitively interpreted. However, it is not guaranteed that, it will always happen. If the observed variables are not highly correlated, at least in groups, then it is likely that FA will not yield satisfactory results.

(3) In PCA, one starts with the maximum number of PCs, and the dimension may be reduced later. Factor extraction must be done with a pre-decided number of factors. Hence, several alternative values of *m* may be considered before a final value of *m* is determined.

(4) PCA aims at explaining variances, FA aims at explaining correlations

(5) PCA is exploratory and does not have any model assumptions, FA is based on a statistical model with assumptions

3.3. Extraction of Latent Factors

Though the relationship between $X_1, X_2, \ldots, X_p$ and f_1, f_1, …, f_m (Sec 3.1) looks like a system of regression equations, estimation of the latent factors needs a different methodology. Two methods are commonly used to extract the factors: Maximum Likelihood Estimation (ML) and Principal Component Extraction. We will employ the latter.

The system of equations in Sec 3.1 may be rewritten as follows

$$\mathbf{f_1} = b_{11}X_1 + b_{12}X_2 + \ldots + b_{1p}X_p$$

$$\mathbf{f_2} = b_{21}X_1 + b_{22}X_2 + \ldots + b_{2p}X_p$$

$$\vdots$$

$$\mathbf{f_j} = b_{j1}X_1 + b_{j2}X_2 + \ldots + b_{jp}X_p$$

$$\vdots$$

$$\mathbf{f_m} = b_{p1}X_1 + b_{p2}X_2 + \ldots + b_{pp}X_p$$

where the common factors are expressed as linear combination of the original variables. In fact, as in case of PCA, we will use the scaled variables only.

The first decision to make regarding factor extraction is to determine *m*. Once *m* is fixed, we proceed to extract the common factors. Since we employ PC method of extraction, we use the knowledge gained in PCA to fix *m*.

Case Study

We use the same "places" data from the Places Rated Almanac. Objective is to identify a set of latent factors, if possible, which will be able to relate to groups of observed variables, and thereby reduce the dimension of the data.

For this problem, the optimum number of PCs is 5. Hence it is decided to extract 5 factors.

[***Note**: Always perform PCA on a data set and decide on the number of principal components. Optimum number of principal components is a subjective choice to some extent (see Sec 2.6.1). Principal components do not necessarily have any interpretability, but factors do. It may help to consider more than one single choice of m, and extract more than one set of factors. Final decision may be taken depending on interpretability*]

3.3.1. Consideration before Factor Extraction

Before actual factor extraction takes place, to make sure that factors are reliable and interpretable, several conditions need to be satisfied. Adequate sample data is required, so that for every factor there are at least 10 observations. If 5 factors are to be extracted, sample size needs to be at least 50.

Following two tests are performed to assess the suitability of the data for factor analysis.

3.3.1.1. Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy

The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy (MSA) is an index used to examine how appropriate FA is. The statistic is a measure of the proportion of variance among variables that might be common variance. Thus, it measures a degree of association among the variables.

KMO returns values between 0 and 1 for each variable in the data set. If KMO of a variable is large then that variable enjoys substantial association with a subset of variables in the data. However, the cut-off may be case dependent. Generally, if MSA is less than 0.5, FA is not recommended, since no reduction is expected. On the other hand, MSA > 0.7 is expected to provide considerable reduction in dimension and extraction of meaningful factors. This is a strong indication of possible existence of an underlying common factor which is manifested through this set. FA will be appropriate in that situation.

On the other hand, a small value of KMO (<0.5) indicate that the corresponding variable does not have strong association with any other variable. Thus, this variable will be associated with one factor by itself. If the overall KMO is small, then no variable is strongly associated with any other variables, and thus FA is not expected to reduce dimension. Each variable in that case would account for a separate factor.

Case Study continued

For factor analysis Python package `factor_analyzer` is to be installed.

Note that we will use scaled variables only. Hence instead of the variance-covariance matrix of the original variables in Places Rated data, the correlation matrix is used.

Recall that high pairwise correlation was observed among population, arts and healthcare (Sec 2.3). For the rest only moderate correlation was observed. But KMO MSA goes beyond the pairwise correlation and takes a more holistic approach to identify overall association.

```
# KMO Test
print("Overall MSA :",np.round(calculate_kmo(std_places)[1],2))
Overall MSA : 0.76

msa = np.round(calculate_kmo(std_places)[0],2)
pd.DataFrame(msa,index=std_places.columns,columns=['MSA']).T
```

	Climate	HousingCost	HlthCare	Crime	Transp	Educ	Arts	Recreat	Econ	Pop
MSA	0.56	0.64	0.85	0.74	0.86	0.79	0.8	0.82	0.39	0.79

Except for economies, all other variables enjoy high MSA. The overall MSA is also substantially high: 0.76. FA is justified in this case.

3.3.1.2. Bartlett's Test of Sphericity

Bartlett's test of sphericity tests the hypothesis that the variables are uncorrelated in the population. In other words, the population correlation matrix is an identity matrix; the diagonal entries are all 1 and off diagonal entries are all 0.

H_0: All variables in the data are uncorrelated

Ha: At least one pair of variables in the data are correlated

If the null hypothesis cannot be rejected, then FA is not advisable, since each variable is associated with a factor and no dimension reduction is possible.

If the p-value is small, then we have sufficient evidence to believe that the correlation matrix is not identity.

Case Study continued

```
# Bartlett test
chi_square_value,p_value=calculate_bartlett_sphericity(std_places)
print('Chi-squared value :',chi_square_value)
print('P-value :',p_value)
Chi-squared value : 1629.9805027725658
P-value : 9.2446407534e-313
```

The very small p-value justifies that performing FA is meaningful for our dataset.

3.3.2. Factor Extraction

Case Study continued

Having made sure that Places Rated data is suitable for factor extraction, we proceed to extract the factors.

```
# Extract 5 Factors
fa = FactorAnalyzer(n_factors=5, rotation=None,method='principal')
factor_no_rotation= fa.fit(std_places)
fa_components= pd.DataFrame(np.round(fa.loadings_,2),columns=  ['pc1',
'pc2','pc3','pc4','pc5'] ,index = std_places.columns)
fa_components['h2'] = np.round(fa.get_communalities(),2)
fa_components['u2']  = np.round(fa.get_uniquenesses(),2)
fa_components
```

	pc1	pc2	pc3	pc4	pc5	h2	u2
Climate	0.36	0.24	-0.75	0.13	0.19	0.80	0.20
HousingCost	0.60	0.38	-0.22	0.51	-0.06	0.82	0.18
HlthCare	0.88	-0.29	0.01	0.05	0.11	0.88	0.12
Crime	0.51	0.37	0.17	-0.57	0.23	0.81	0.19
Transp	0.62	-0.09	0.18	-0.09	-0.59	0.78	0.22
Educ	0.50	-0.40	0.29	0.45	0.01	0.69	0.31
Arts	0.90	-0.20	-0.02	-0.10	0.14	0.88	0.12
Recreat	0.57	0.47	-0.07	-0.13	-0.44	0.76	0.24
Econ	0.21	0.59	0.62	0.28	0.27	0.93	0.07
Pop	0.87	-0.23	-0.02	-0.19	0.21	0.88	0.12

```
fa_variance =
pd.DataFrame(np.round(fa.get_factor_variance(),2),columns=['fa1','fa2','fa3','fa4','fa5'],index
= ['SS Loadings','Proportion Var','Cumulative Var'])
fa_variance
```

	fa1	fa2	fa3	fa4	fa5
SS Loadings	4.09	1.26	1.14	0.95	0.79
Proportion Var	0.41	0.13	0.11	0.10	0.08
Cumulative Var	0.41	0.54	0.65	0.74	0.82

PC1 – PC5 are the factors extracted using principal component method. They happen to be identical to the PCs found in Sec 2.5 except possibly for opposite sign in a few cases, which is equivalent to multiplying a PC by (- 1). That does not change the essential properties of the PCs. Hence the extracted PCs are also equally acceptable. Recall that the coefficients are identical to the correlation coefficients between the original variables and PCs as shown in Sec 2.5, except for the signs corresponding to a few columns. These coefficients are called **factor loadings**.

Each variable (scaled) can be expressed as a linear function of the 5 factors. For example

SClimate = 0.36 f_1 + 0.24 f_2 – 0.75 f_3 + 0.13 f_4+ 0.19 f_5 + ε_1

SHousingCost = 0.60 f_1 + 0.38 f_2 – 0.22 f_3 + 0.51 f_4 -0.06 f_5 + ε_2

The column h2 is the communality. Recall that the definition of communality is the sum of square of the coefficients of the factors in the expression of (scaled) variables. Communality of Climate is 0.80. That means 80% of variance in climate is explained by the common factors. Since the variables are all scaled, each has variance is 1. The extra variance that is not explained by the common factors is due to the errors or the **specific variance**. For climate that value is 20% and is given under the column u2.

Except for education, values of specific variances are all small. That indicates the 5-factor model works adequately for the Places Rated data.

The factor variances are given in SS loadings. Note that here the variances are identical to the variances of the PCs (Table 1) since the factors are identical to the PCs themselves. Total variance in the data is 10, equal to the number of variables in the data. *Proportion Var* and *Cumulative Var* compute the proportion of variance and cumulative variance corresponding to the factors, respectively. With 5 factors 82% of the total variance in the data is explained.

3.3.3. Factor Rotation

Recall that the latent factors are expected to explain groups of observed variables. That means factors are expected to show high correlation with a subset of the variables and no correlation with the complementary set of variables. In practice, that does not always happen.

In Places Rated Data Factor 1 (PC1) loads heavily on healthcare, arts and population. It also shows moderate correlation with housing cost and transport. We may name this factor Urbanity, since this factor is expected to have high scores for large metropolitan areas. On the other hand, the other factors (PC2 – PC5) do not seem to load heavily on any variable. So it is difficult to interpret these 4 factors.

In order to improve interpretability, the factors are often rotated. Usually the rotations are orthogonal, but oblique rotations are also possible. The most common rotation is varimax which is expected to increase the already large correlations, and reduce the already small correlations. Rotation does not affect communality; neither does it change the total variance explained by the 5 factors together.

```
# Factor rotation
fit1 = FactorAnalyzer(n_factors=5, rotation='varimax', method='principal')
factor_rotation= fit1.fit(std_places)
fit1_components= pd.DataFrame(fit1.loadings_,columns=  ['Rc1', 'Rc2','Rc3','Rc4','Rc5'] ,index
= std_places.columns)
fit1_components['h2']  = np.round(fit1.get_communalities(),2)
fit1_components['u2']  = np.round(fit1.get_uniquenesses(),2)
np.round(fit1_components,2)
```

	Rc1	Rc2	Rc3	Rc4	Rc5	h2	u2
Climate	0.12	-0.02	0.86	-0.16	0.12	0.80	0.20
HousingCost	0.27	0.35	0.67	0.38	-0.15	0.82	0.18
HlthCare	0.89	0.21	0.18	0.03	0.08	0.88	0.12
Crime	0.26	0.18	0.03	0.20	0.81	0.81	0.19
Transp	0.41	0.77	-0.11	-0.04	-0.02	0.78	0.22
Educ	0.68	0.09	-0.08	0.22	-0.41	0.69	0.31
Arts	0.85	0.23	0.20	0.00	0.25	0.88	0.12
Recreat	0.08	0.75	0.31	0.14	0.28	0.76	0.24
Econ	0.02	0.04	-0.03	0.95	0.15	0.93	0.07
Pop	0.85	0.15	0.17	-0.04	0.33	0.88	0.12

```
fit1_variance= pd.DataFrame(fit1.get_factor_variance(),columns=['Rc1',
'Rc2','Rc3','Rc4','Rc5'] ,index = ['SS Loadings', 'Proportion Var','Cumulative Var'])
np.round(fit1_variance,2)
```

	Rc1	Rc2	Rc3	Rc4	Rc5
SS Loadings	3.04	1.44	1.42	1.19	1.15
Proportion Var	0.30	0.14	0.14	0.12	0.11
Cumulative Var	0.30	0.45	0.59	0.71	0.82

RC1, RC2, RC3, RC4 and RC5 are the 5 factors obtained after rotation.

Note that RC1 gives maximum loadings to healthcare, arts and population so that the Urbanity factor is maintained. RC2 has highest loading with economies and small loading with the rest. RC2 therefore may be named Personal Economics. RC3 has highest loading with climate and housing cost. We may name this factor Pleasant Living. RC4 shows high loading with crime and the loading is positive, indicating the higher the score of a metro on this factor, the worse it will be for living. Let us call this factor Crime. The highest loading for RC5 is with transport and recreation. It may

be difficult to interpret a single factor that would explain these two attributes together. In absence of a common quality, we name factor 5 Outdoors.

All variables except education is thus loaded to one factor and one factor only. Education has the lowest communality, indicating that the proportion of its variance explained by all factors is low. The highest loading of education is with Urbanity. Therefore, education has also been assigned to Factor 1.

Latent factors	Name of factor	Variable names
Factor 1	Urbanity	healthcare, arts, education, population
Factor 2	Personal Economics	economics
Factor 3	Pleasant Living	climate, housing cost
Factor 4	Crime	crime
Factor 5	Outdoors	transport, recreation

Table 3: Latent Factors loaded on Actual Variables

3.3.4. Factor Scores and their Usage

Similar to the PC scores, Factor scores are also obtained by multiplying the scaled observations by the factor loadings. Each and every observation has a score value corresponding to each factor. The full table is given in Appendix 1.

The most important practical usage of extracting the latent factors is to cluster and rank observations according to various criteria. While taking a decision on living or working in a certain metropolitan area, consumers may have different requirements. Instead of making a choice based on several related attributes, it is always easier to choose based on a single composite ranking that is available through factor scores.

Consider first the factor called **Urbanity**. The top rank holders are the largest cities in US, such as New York and Los Angeles. Though Washington D.C and Boston are not as large as these metabolizes, by Urbanity index they are comparable.

```
# Top 5 metros by Urbanity Score
```

```
scores = pd.DataFrame(fit1.transform(std_places),columns=['Rc1', 'Rc2','Rc3','Rc4','Rc5'])
scores['names'] = names
scores = scores.sort_values(by=['Rc1'],ascending=False)
np.round(scores[['Rc1','names']].head(6),2)
```

	Rc1	names
212	7.86	New-York,NY
64	5.19	Chicago,IL
233	4.62	Philadelphia,PA-NJ
178	4.60	Los-Angeles,Long-Beach,CA
313	3.93	Washington,DC-MD-VA
42	3.49	Boston,MA

The second factor Personal Economics outlook should have higher ranking for the metro areas where cost of living is reasonable. The top areas determined by the highest score in this factor are the metros which have cost of living index at par or below the country average.

```
scores = scores.sort_values(by=['Rc4'],ascending=False)
np.round(scores[['Rc4','names']].head(6),2)
```

	Rc4	names
194	3.20	Midland,TX
159	3.04	Lafayette,LA
240	2.80	Portsmouth-Dover-Rochester,NH-ME
20	2.67	Atlantic-City,NJ
317	2.55	West-Palm-Beach-Boca-Raton-Delray-Beach,FL
109	2.53	Fort-Pierce,FL

Similarly, the factor Pleasant Living is mostly associated with moderate climatic conditions, as a result of which all of these areas have high housing cost. No wonder that many of the metros are in California.

```
# Top 5 metros by Pleasant Living Score
scores = scores.sort_values(by=['Rc3'],ascending=False)
np.round(scores[['Rc3','names']].head(6),2)
```

	Rc3	names
226	3.89	Oxnard-Ventura,CA
289	3.88	Stamford,CT
10	3.72	Anaheim-Santa-Ana,CA
268	3.28	San-Diego,CA
271	3.22	Santa-Barbara-Santa-Maria-Lompoc,CA
216	3.20	Norwalk,CT

Factor 4 is all about Crime and only crime. Several large cities, such as New York and Los Angeles rank high on this index. Along with them are several other small to mid-sized metro area, such as Odessa and Las Vegas, which also enjoy notoriety.

```
# Top 5 metros by Crime Score
scores = scores.sort_values(by=['Rc5'],ascending=False)
np.round(scores[['Rc5','names']].head(6),2)
```

	Rc5	names
212	5.23	New-York,NY
178	3.36	Los-Angeles,Long-Beach,CA
191	3.32	Miami-Hialeah,FL
167	2.75	Las-Vegas,NV
219	2.02	Odessa,TX
85	1.90	Detroit,MI

In this instance it would be interesting in seeing which metro areas have lowest score in crime. Table 3 below shows the bottom 5 areas.

Serial No.	Metropolitan Areas	Crime
290	Stamford,CT	-2.40
217	Norwalk,CT	-2.36
55	Burlington,VT	-2.17
76	Cumberland,MD-WV	-2.06
193	Middlesex-Somerset,Hunterdon,NJ	-1.94

Table 4: Metro areas with lowest score in Crime

Factor 5 Outdoors gives maximum loadings to transportation and recreation. The places with good opportunities for recreational activities like good restaurants, public golf courses, certified lanes for tenpin bowling, movie theatres, zoos, etc. have been given high score in the original data. Incidentally, the ability to make use of these recreational opportunities is closely related to good transportation facilities. So, it is intuitive that Outdoors can explain both transportation and recreation. Also, the places with high scores in transportation are the major cities thriving with economic activities where public transport is easily available.

```
# Top 5 metros by Outdoors Score
scores = scores.sort_values(by=['Rc2'],ascending=False)
np.round(scores[['Rc2','names']].head(6),2)
```

	Rc2	names
43	3.39	Boulder-Longmont,CO
269	3.27	San-Francisco,CA
54	2.84	Burlington,VT
277	2.71	Seattle,WA
265	2.69	Salt-Lake-City-Ogden,UT
167	2.62	Las-Vegas,NV

We can see that the area with the best score is Boulder-Longmont, Colorado. It is an area with an extensive public bus system with cycling opportunities. Residents can enjoy a plethora of outdoor activities like rock climbing, plunging into a freezing reservoir, major running events, and cycling.

4. Appendix

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

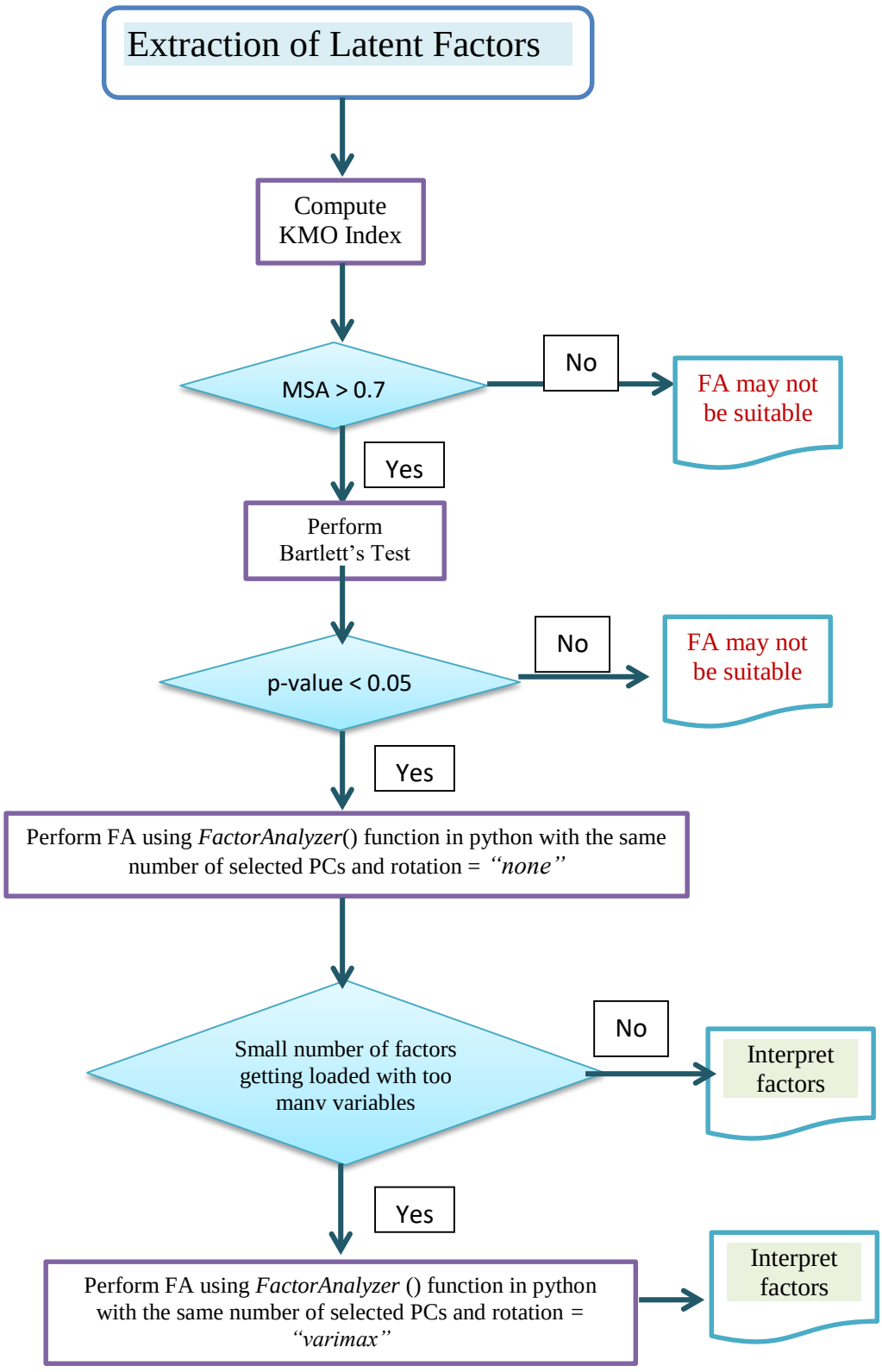


Fig 3: Flowchart for Factor Analysis

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/