FACTOR ANALYSIS

Introduction

- Factor Analysis is similar to PCA in that it is a technique for studying the interrelationships among variables.
- Both methods differ from regression in that they don't have a dependent variable.
- A goal in PCA and Factor Analysis is to obtain a new set of distinct summary variables, which are fewer in number than the original number of variables.
- The goal of PCA is to select a new set of uncorrelated variables (principal components) that explain as much of the total variation as possible.
- In Factor Analysis the goal is to identify factors that explain the interrelationships among the original variable.
    - Our goal is to remove redundancy or duplication from a set of correlated variables.
- The assumption behind Factor Analysis is that there is a number of ***latent variables*** or *factors* that account for the correlations among the ***original variables***, such that if the factors are held constant, the partial correlations among the observed variables all become zero.
- Therefore, the latent factors determine the values of the observed variables.
- Factor Analysis has been used in many of the same research areas as PCA (e.g. psychometrics, chemometrics, and econometrics).
- Overview of applications of Factor Analysis
    a. Identification of Underlying Factors
        i. Clusters variables into homogenous sets
        ii. Creates new variables (i.e. factors)
    b. Screening of variables
        i. Allows us to select one variable (factor) to represent many.

c. Clustering of objects

    i. Allows us to put objects (people) into categories depending on their factor scores.

- Factor Analysis has been used less than PCA likely because it has many options during the analysis and the results may not be as straight forward as those for PCA.

> "Perhaps the most widely used (and misused) multivariate [technique] is factor analysis. Few statisticians are neutral about this technique. Proponents feel that factor analysis is the greatest invention since the double bed, while its detractors feel it is a useless procedure that can be used to support nearly any desired interpretation of the data. **The truth, as is usually the case, lies somewhere in between**. Used properly, factor analysis can yield much useful information; when applied blindly, without regard for its limitations, it is about as useful and informative as Tarot cards. **In particular, factor analysis can be used to explore the data for patterns, confirm our hypotheses, or reduce the Many variables to a more manageable number."**
>
> --Norman Streiner, *PDQ Statistics*

http://ocw.jhsph.edu/courses/statisticspsychosocialresearch/pdfs/lecture8.pdf

Factor Analysis vs. Principal Component Analysis

- The variance of a variable consists of two components,
    - A *common* variance that is shared with other variables in the model, and
    - A *unique* variance that is specific to a variable and includes an error component.
- A *common factor* is an unobservable, hypothetical variable that contributes to the variance of at least two of the observed variables.
- A *unique factor* is an unobservable, hypothetical variable that contributes to the variance of only one observed variable.

Table 1. Differences between Principal Component Analysis and Factor Analysis.

	Principal component analysis	Factor analysis
Variance considered	Considers the total variance and makes no distinction between the common and unique variance	Considers only the common variance of the variables
Purpose of the method	1. To account for the maximum portion of the variance with a minimum number of new or composite variables called principal components.	1. To extract a smaller number of factors that account for the intercorrelations among the observed variables and 2. Identify the latent dimensions that explain why the observed variables are related.
When method is appropriate	Variables are measured with little error; thus, the error and specific variance represent a small portion of the total variance in the original variables.	Observed variables are only latent constructs to be measured or if the error variance makes up a significant portion of the total variance

Basic Concepts

- Suppose we have a set of variables $X_1, X_2, \ldots X_P$
    - To simplify the description of these variables, we will subtract the mean of each dataset from each observation; thus,
        - $x_1 = (X_1 - \bar{X}_1)\,, x_2 = (X_2 - \bar{X}_2), \ldots, x_P = (X_P - \bar{X}_P)$
- The purpose of Factor Analysis is to represent each of these variables as a linear combination of a smaller set of common factors plus a factor unique to each of the response variables.
- These linear combinations can be represented by:

$$x_1 = l_{11}F_1 + l_{12}F_2 + \ldots + l_mF_m + e_1$$

$$x_2 = l_{21}F_1 + l_{22}F_2 + \ldots + l_mF_m + e_2$$

.

.

$$x_P = l_{P1}F_1 + l_{P2}F_2 + \ldots + l_{Pm}F_{Pm} + e_1$$

where:

1. m is the number of common factors, which is typically much smaller than the number of parameters (P).

2. $F_1, F_2, \ldots F_m$ are the common variances.

3. l_{ij} is the coefficient of F_j in the linear combination describing x_i. This term is referred to as the ***loading*** of the i^{th} variable of the j^{th} factor.

4. $e_1, e_2, \ldots, e_P$ are the *unique factors*, each relating to one of the original variables.

- Let's assume $x_1, x_2, x_3, x_4,$ and x_5 represent values on five different traits associated with malt quality of a line and ***m=2***, the model would be:

$$x_1 = l_{11}F_1 + l_{12}F_2 + e_1$$

$$x_2 = l_{21}F_1 + l_{22}F_2 + e_2$$

$$x_3 = l_{31}F_1 + l_{32}F_2 + e_3$$

$$x_4 = l_{41}F_1 + l_{42}F_2 + e_4$$

$$x_5 = l_{51}F_1 + l_{52}F_2 + e_5$$

  - Each of the five values consist of two parts
    - Part due to the common Factors F_1 and F_2
    - Part due to the unique factors e_i.
  - The factor analysis is similar to the PCA in that each Principal Component is expressed as a linear combination of the variables. The number of principal components is equal to the number of variables.
  - In Factor Analysis, we choose a number of factors that is smaller than the number of independent variables.
    - Often times we know the number of Factors (m) variables before the analysis.
    - If you don't know m, there are ways to determine the number of factors.

- In the models above we not only broke each independent variable into the common factors (F_i) and unique factors (e_i), we can break down the variance of x_i into two parts, the ***communality*** ($\boldsymbol{h_i^2}$) and the ***specificity*** ($\boldsymbol{u_i^2}$).
    - Communality: Part of the variance that due to the common factors.
    - Specificity: Part of the variance that is due to the unique factor.
    - Thus, the variance of $x_i = 1 = h_i^2 + u_i^2$
- The numerical answer we most often look for from Factor Analysis are:
    - Estimates of the **Factor Loadings (l_{ij})**
    - Estimates of the **Communalities** $\boldsymbol{(h_i^2)}$.

Assumptions of the Factor Analysis model

1. Measurement errors are homogenous and an average value of 0.
2. There is no association between the factor and the measurement's error
    a. Cov(F, e_j)=0
3. There are no associations between errors.
    a. Cov(e_i, e_j)=0
4. Given the factor, there is no association between the observed variables .
    a. Cov(X_j, $X_k|F$)=0

Performing the Factor Analysis

- Any Factor Analysis typically includes three main steps. Items highlighted in bold font will be the methods we discuss in class.

1. Initial extraction of Factors
    a. Principal component method (default method in SAS)
    b. Principal factor analysis (iterated approach)
    c. Maximum likelihood

2. Factor rotations

    a. **Varimax rotation**

    b. Oblique rotation

    c. Direct quartimin rotation

3. Assigning Factor Scores

Principal Component Analysis for Initial Extraction

- The basis of the use of PCA for initial extraction is to choose the first *m* principal components and modify them to fit the factor model.

- The reason for choosing the first *m* principal components is that they explain the most variation and are thus considered the "most" important.

- Remember from PCA that the principal components are:

$$C_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1P}x_P$$

$$C_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2P}x_P$$

$$\vdots$$

$$C_P = a_{1P}x_1 + a_{2P}x_2 + \ldots + a_{PP}x_P$$

- It can be shown mathematically that these equations can be inverted to express the x_i*'s* as functions of the principal components (C_j*'s*). The result is:

$$x_1 = a_{11}C_1 + a_{21}C_2 + \ldots + a_{P1}C_P$$

$$x_2 = a_{12}C_1 + a_{22}C_2 + \ldots + a_{P2}x_P$$

$$\vdots$$

$$x_P = a_{1P}C_1 + a_{2P}C_2 + \ldots + a_{PP}C_P$$

- With a few additional mathematical functions, the each variable x_1 can be expressed in the form:

$$x_i = l_{i1}F_1 + l_{i2}F_2 + \ldots + l_{im}F_{im} + e_i$$

Example of Results from the Initial Extraction Using the Principal Component Analysis

- We will assume that we have five variables ($x_1 - x_5$) and that previous research has determined that there are two factors (F_1 and F_2).

- A summary of that initial analysis is provided in the table below.

	Factor loadings		Communality	Specificity
Variable	F_1	F_2	h_i^2	u_i^2
x_1	0.470	0.734	0.759	0.241
x_2	0.510	0.704	0.756	0.244
x_3	0.481	-0.258	0.298	0.702
x_4	0.888	-0.402	0.949	0.051
x_5	0.956	-0.233	0.968	0.032
Variance explained	2.413	1.317	$\Sigma h_i^2 = 3.730$	$\Sigma u_i^2 = 1.270$
Percentage	48.3	26.3	74.6	25.4

- Based on these results, the **loading** of x_1 on F_1 is $l_{11} = 0.470$ and on F_2 is l_{12} is 0.734.

- The equation for the factor model for x_1 is $x_1 = 0.470F_1 + 0.734F_2 + e_1$

- The total variance explained by each factor is 2.413 (48.3%) for F_1 and 1.317 (26.3%) for F_2.

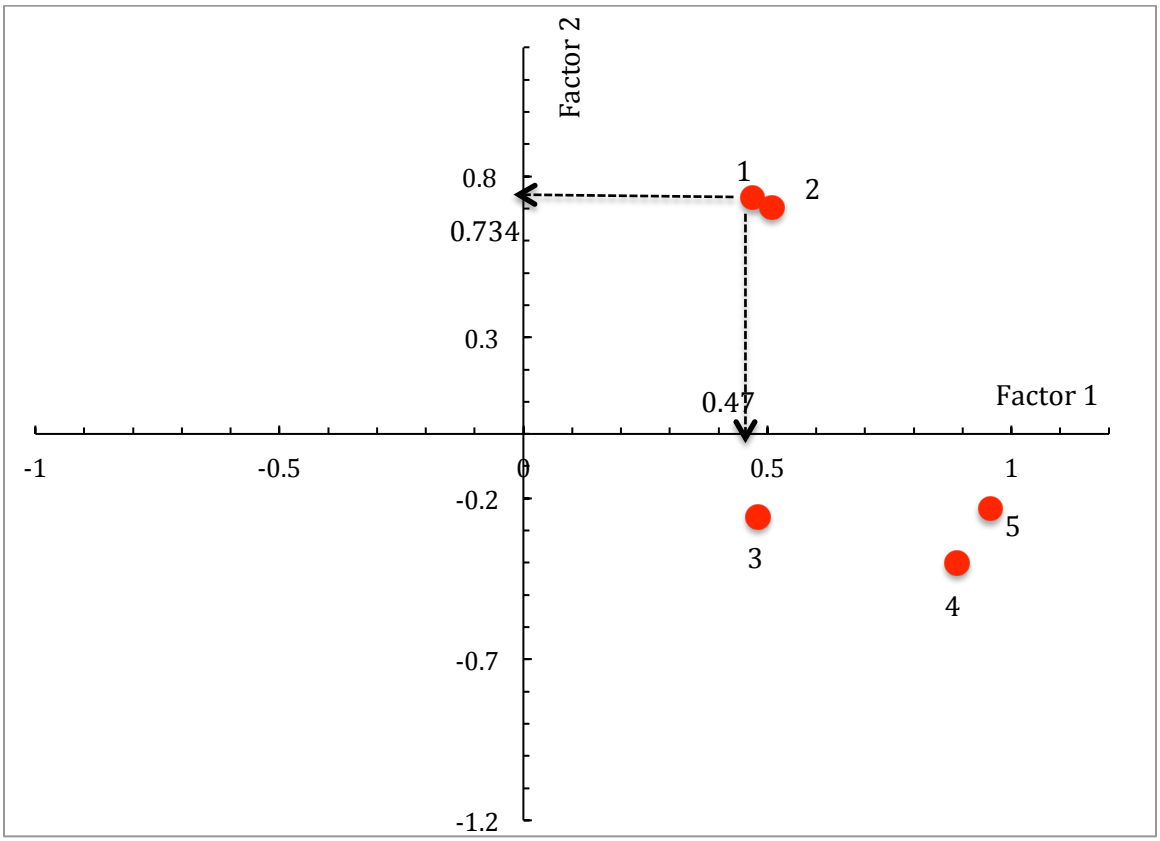
- The communality value (h_i^2) shown for each variable shows the part of the variance of each variables explained by the **common factors**.

    - e.g. for x_1 $h_i^2 = 0.470^2 + 0.734^2 = 0.759$

- The specificity value (μ_i^2) shown for each variable is the part of the variance not explained by h_i^2.

    - $\mu_i^2 = 1 - h_i^2$

- A valuable tool for interpreting the output from Factor Analysis are the **Factor Diagrams**.



Factor Rotations

- Remember the purpose of factor analysis is to derive from the data *"easily interpretable common factors."*
- However, the results from the initial analysis are difficult to interpret.
- More easily interpretable factors, call **Rotated Factors**, can be obtained through a process called factor ration.
- The goal of factor ration is to obtain some of the loadings that are very large (near ±1) and the remaining loadings are very small (near 0).
- Additionally, for any given variable we want it to have a high loading for only **one factor**.

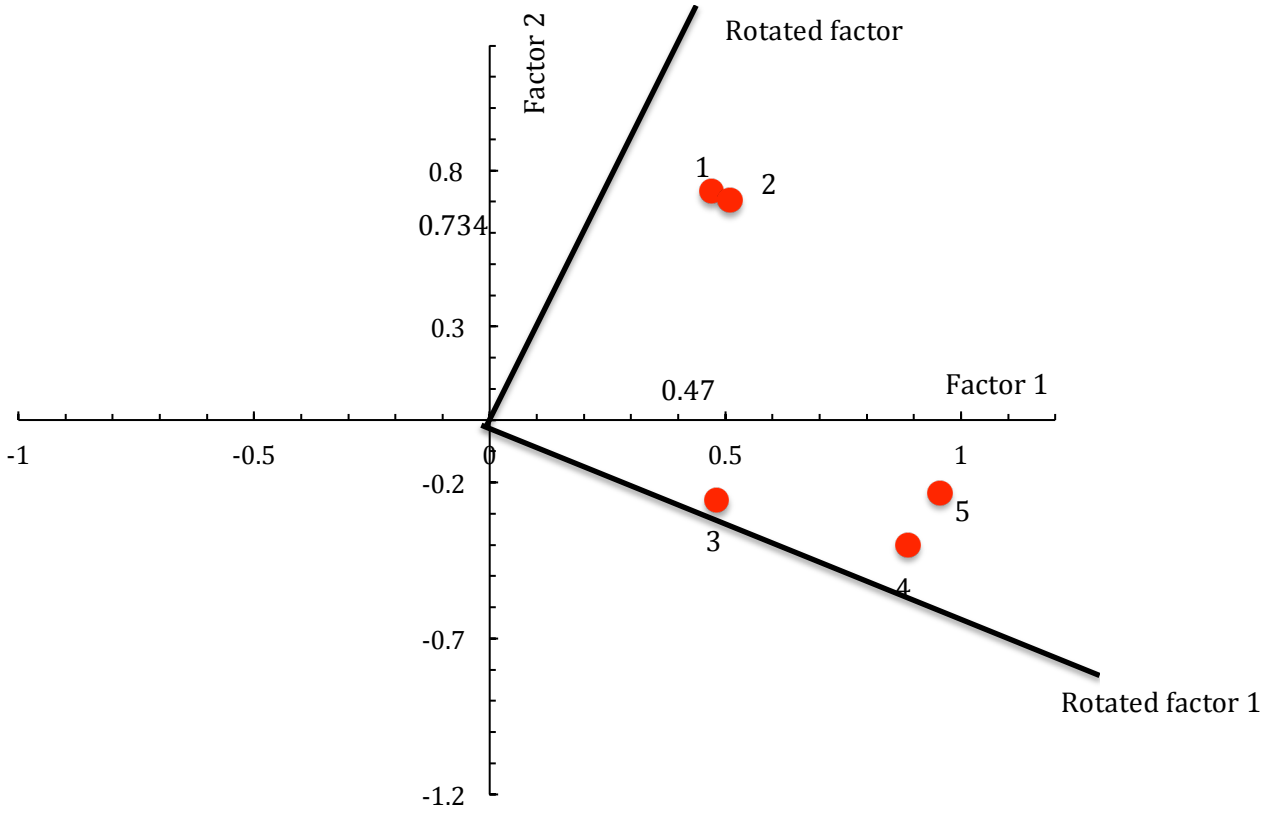
- This will allow us to give each factor an interpretation that comes from the variables with which it is highly correlated.

- In class we will discuss only the most common method of factor ration, the **Varimax Rotation**.

- Any new rotation results in "new axes" that go through or near clusters of responses that represent the response variable.

- In the Varimax Rotation, the new axes must be perpendicular to each other. This makes axes orthogonal or independent of each other.

- The figure below shows the Varimax Rotation where the axis for Factor 1 was rotated down to be near the variables x_3, x_4, and x_5; and Factor 2 was rotated to the right to be near variables x_1 and x_2.



- The results show that variable x_1 and x_2 have high loadings on rotated Factor 2 and variables x_3, x_4, and x_5 have high loadings on rotated variable 1.

- The table below shows the new loadings based on the Varimax Rotation.
    - Note that the communality values before and after the rotation were not changed.

Variable	Factor loadings		Communality
	F_1	F_2	h_i^2
x_1	0.063	**0.869**	0.759
x_2	0.112	**0.862**	0.756
x_3	**0.546**	0003	0.598
x_4	**0.972**	0.070	0.949
x_5	**0.951**	0.251	0.968
Variance explained	2.164	1.566	$\Sigma h_i^2 = 3.730$
Percentage	43.3	31.3	74.6

Assigning Factor Scores

- It may be a goal of your research to determine the value each individual has for each factor.
- The simplest was is to add together the loads of the respective.
    - For example, to determine the value for Factor 1 for each variable, you would add together the values for x_3, x_4, and x_5 *=(0.546+0.972+0.951) = 2.469*.
    - To determine the value for Factor 2 for each variable, you would add together the values for x_1 and x_2 *= (0.869+0.862) = 1.731*.
    - There are many more ways to assign factor scores than this simple method described above, including methods that involve regression. You should consult with a statistician if you want to consider a more complicated method of reporting factor scores.

SAS Example

- This example analyzes socioeconomic data from 12 different neighborhoods in the Los Angeles area. The five variables represent total population (*Population*), median school years (*School*), total employment (*Employment*), miscellaneous professional services (*Services*), and median house value (*HouseValue*). **Our goal in the factor analysis is to determine if like variables can be grouped into factors and to determine a name that describes the type of data in each factor.**

- SAS Commands

```
data SocioEconomics;
input   Population      School Employment   Services        HouseValue;
  datalines;
5700    12.8    2500    270     25000
1000    10.9    600     10      10000
3400    8.8     1000    10      9000
3800    13.6    1700    140     25000
4000    12.8    1600    140     25000
8200    8.3     2600    60      12000
1200    11.4    400     10      16000
9100    11.5    3300    60      14000
9900    12.5    3400    180     18000
9600    13.7    3600    390     25000
9600    9.6     3300    80      12000
9400    11.4    4000    100     13000
;
ods graphics on;
ods rtf file='factor.rtf';
proc corr;
title 'Simple Linear Correlation Results';
run;
proc factor data=SocioEconomics
        rotate=varimax
        outstat=fact_all
        plots=(scree initloadings preloadings loadings);
title 'Factor Analysis';
run;
ods graphics off;
ods rtf close;
```

Simple Linear Correlation Results

The CORR Procedure

5 Variables:	Population School Employment Services HouseValue

Simple Statistics						
Variable	**N**	**Mean**	**Std Dev**	**Sum**	**Minimum**	**Maximum**
Population	12	6242	3440	74900	1000	9900
School	12	11.44167	1.78654	137.30000	8.30000	13.70000
Employment	12	2333	1241	28000	400.00000	4000
Services	12	120.83333	114.92751	1450	10.00000	390.00000
HouseValue	12	17000	6368	204000	9000	25000

Pearson Correlation Coefficients, N = 12 Prob > |r| under H0: Rho=0					
	Population	**School**	**Employment**	**Services**	**HouseValue**
Population	1.00000	0.00975 0.9760	0.97245 <.0001	0.43887 0.1535	0.02241 0.9449
School	0.00975 0.9760	1.00000	0.15428 0.6321	0.69141 0.0128	0.86307 0.0003
Employment	0.97245 <.0001	0.15428 0.6321	1.00000	0.51472 0.0868	0.12193 0.7058
Services	0.43887 0.1535	0.69141 0.0128	0.51472 0.0868	1.00000	0.77765 0.0029
HouseValue	0.02241 0.9449	0.86307 0.0003	0.12193 0.7058	0.77765 0.0029	1.00000

We will see in the Factor Analysis results that these two groups will be divided between the two Factors in a similar manner.

Example of Factor Analysis

The FACTOR Procedure

Input Data Type	Raw Data
Number of Records Read	12
Number of Records Used	12
N for Significance Tests	12

Example of Factor Analysis

The FACTOR Procedure
Initial Factor Method: Principal Components

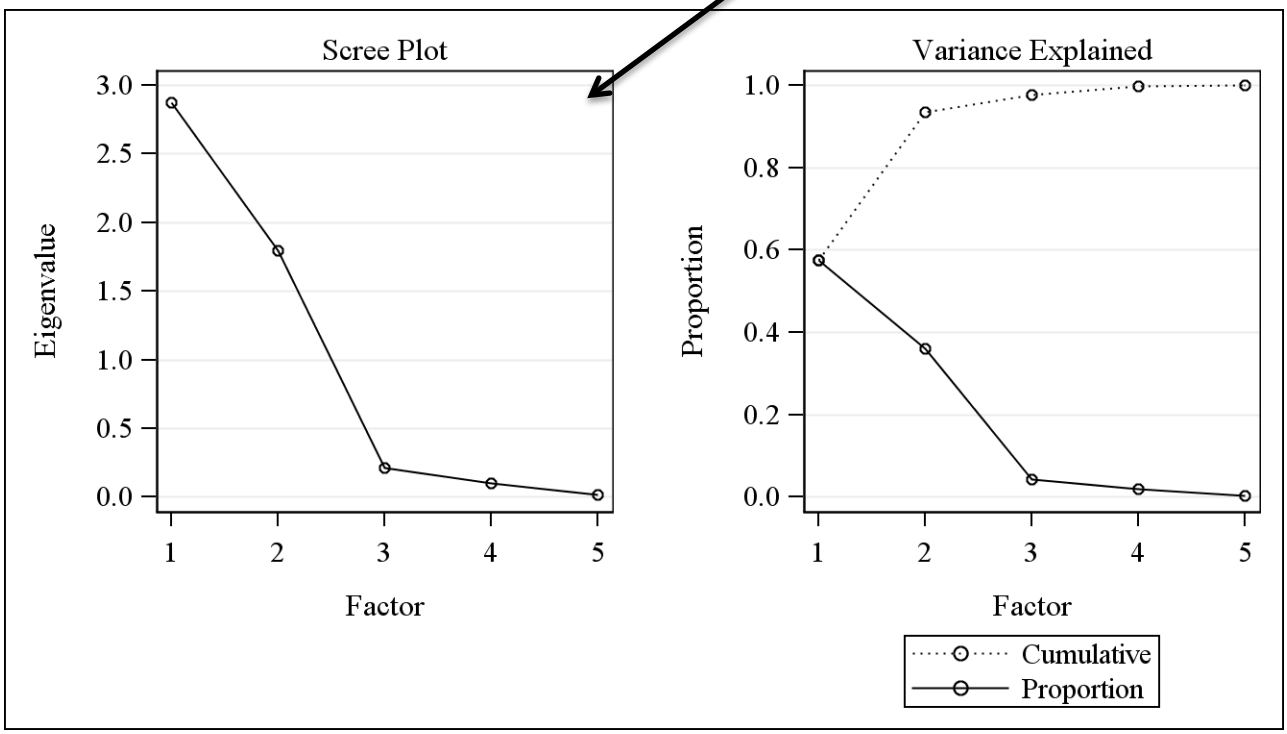
Prior Communality Estimates: ONE

By default, SAS determined the number of factors (*m*) based on the number of Eigenvalues >1.0.

Eigenvalues of the Correlation Matrix: Total = 5 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.87331359	1.07665350	0.5747	0.5747
2	1.79666009	1.58182321	0.3593	0.9340
3	0.21483689	0.11490283	0.0430	0.9770
4	0.09993405	0.08467868	0.0200	0.9969
5	0.01525537		0.0031	1.0000

2 factors will be retained by the MINEIGEN criterion.

The results shown in the scree plot agree that the number of factors should be 2. There is a sharp dropoff stops between 2 and 3 factors; thus, *m* should be 2.



Example of Factor Analysis

The FACTOR Procedure
Initial Factor Method: Principal Components

Factor Pattern		
	Factor1	**Factor2**
Population	0.58096	0.80642
School	0.76704	-0.54476
Employment	0.67243	0.72605
Services	0.93239	-0.10431
HouseValue	0.79116	-0.55818

The loading of the variable population on Factor 1 is 0.58096 and the loading of population on Factor 2 is 0.80642.

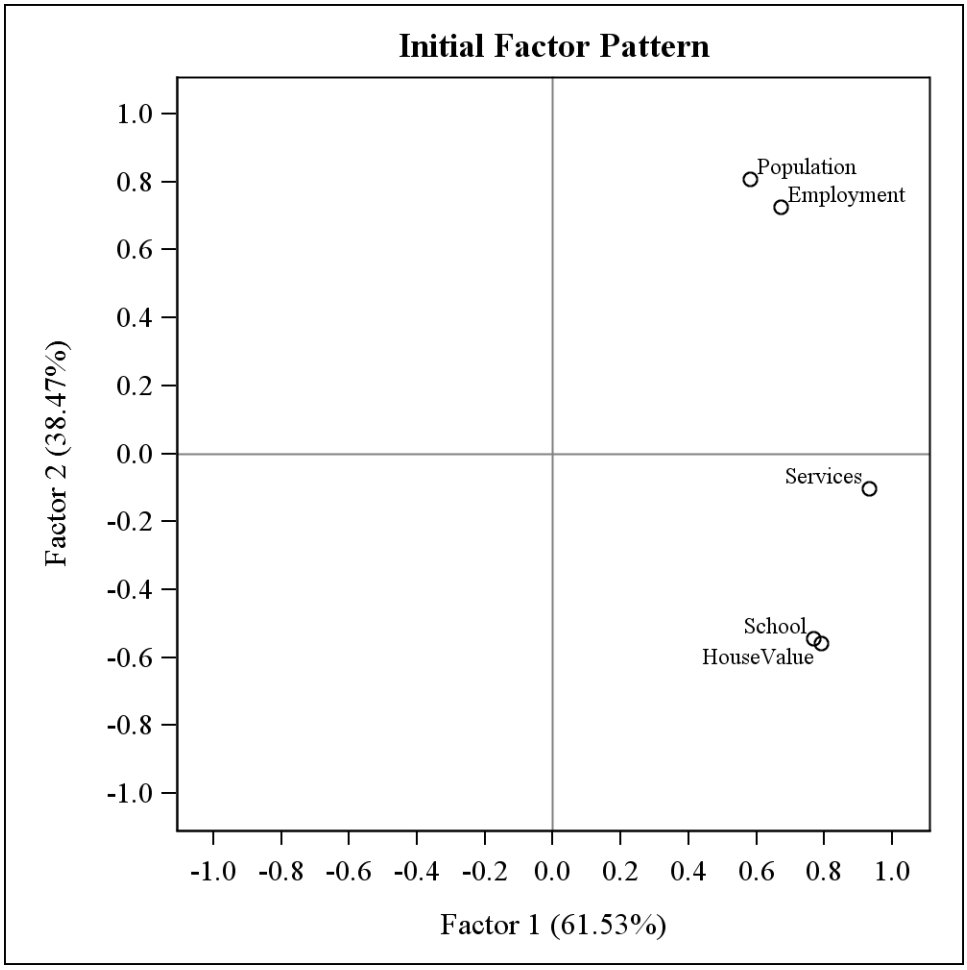
Variance Explained by Each Factor	
Factor1	**Factor2**
2.8733136	1.7966601

Final Communality Estimates: Total = 4.669974				
Population	**School**	**Employment**	**Services**	**HouseValue**
0.98782629	0.88510555	0.97930583	0.88023562	0.93750041

The sum of the Communality Estimates (4.669974) equals the sum of the variances explained by each factor.

Example of Factor Analysis

The FACTOR Procedure
Initial Factor Method: Principal Components



Example of Factor Analysis

The FACTOR Procedure
Rotation Method:
Varimax

Orthogonal Transformation Matrix		
	1	2
1	0.82069	0.57137
2	-0.57137	0.82069

Rotated Factor Pattern		
	Factor1	Factor2
Population	0.01602	0.99377
School	0.94076	-0.00882
Employment	0.13702	0.98007
Services	0.82481	0.44714
HouseValue	0.96823	-0.00605

The loadings of School, Services, and Home Value approach 1.0 for Factor 1.

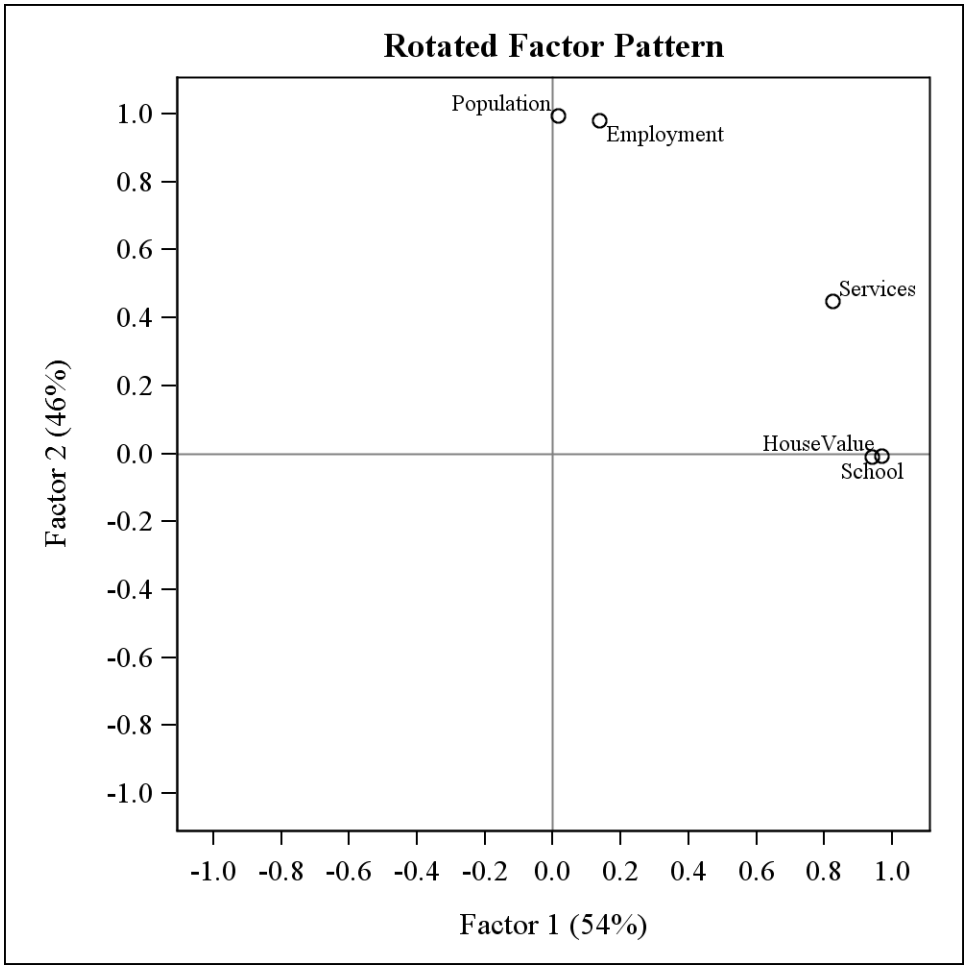
The loading of Population and Employment approach 1.0 for Factor 2.

Variance Explained by Each Factor	
Factor1	Factor2
2.5218278	2.1481459

Final Communality Estimates: Total = 4.669974				
Population	School	Employment	Services	HouseValue
0.98782629	0.88510555	0.97930583	0.88023562	0.93750041

Example of Factor Analysis

The FACTOR Procedure
Rotation Method:
Varimax



In the Varimax Rotation, the axis for Factor 1 was rotated to be near the variables Home Value, School, and Services.

The axis for Factor 2 was rotated so it was near the variables Population and Employment.

SAS Example #2 – Malt Quality Dataset

- SAS commands

```
options pageno=1;
data malt;
input Plump Protein Extract amylase  DP kolbach Solprot Color FAN Betagluc Viscosity
Fructose Glucose Maltose Maltotriose;
datalines;
93.75   12.8    77.9    79.2    152.9   47.35   6.06    2.15    309.4   214.4   1.465   0.0945
        1.05    4.105   0.99
95.1    12.7    76.95   78.95   142.4   50.15   6.36    2.4     380.95  261.95  1.465   0.21
        1.08    3.895   0.97
95.75   13      77.35   76.95   152.7   48.15   6.24    2.4     318.75  202.05  1.45    0.127
        1.04    3.765   0.955
94.2    12.55   78.25   59.65   143.05  45.1    5.665   2.35    361.1   187.6   1.45    0
        0.965   3.97    0.92
                                        .
                                        .
                                        .
                                        .
92.8    11.75   78.6    67      136.9   52.5    6.16    2.9     323.55  173.75  1.455   0
        1.18    4.005   0.97
93.45   11.8    78.25   67.95   146.25  40.6    4.835   2.25    281.5   165.4   1.495   0
        1.165   4.13    0.95
95.3    12.45   77.35   71.4    176.2   45.75   5.685   2       291.8   256.35  1.515   0
        0.975   4.26    0.905
;;
ods graphics on;
ods rtf file='malt.rtf';
proc factor data=malt (drop=fan color fructose) nfactors=4
        rotate=varimax
        outstat=fact_all
        plots=(scree);
run;
ods graphics off;
ods rtf close;
```

From previous research I know there should be around 3-4 factors, so I set the number of factors in the problem using the statement nfactors=4.

In previous Factor Analyses not shown, I found that the variables fan, color, and fructose did not have suitably high loadings under any factor; thus, I decided to drop them from the analysis using the statement (drop=fan color fructose).

Annotated Factor Analysis Results of Malt Quality

The FACTOR Procedure
Initial Factor Method: Principal Components

Rotated Factor Pattern				
	Factor1	**Factor2**	**Factor3**	**Factor4**
Plump	0.07015	0.09034	0.75794	-0.02017
Protein	-0.14630	-0.12427	0.74034	-0.21885
Extract	-0.14320	0.07922	-0.19461	0.79213
amylase	0.26145	0.51283	0.41304	0.27047
DP	0.74436	-0.27137	0.23601	-0.05951
kolbach	-0.19988	0.90558	-0.20511	0.00248
Solprot	-0.26440	0.85055	0.09226	-0.09086
Betagluc	0.24412	0.00193	0.39887	-0.63790
Viscosity	-0.18566	0.16246	-0.39213	-0.63018
Glucose	-0.46115	0.30579	0.01205	0.35885
Maltose	0.78462	0.01755	-0.30906	-0.03710
Maltotriose	-0.78335	0.09846	-0.10501	0.04035

The highlighted values represent the highest loads for each variable.

From previous work, I know the four Factors do represent general trait classifications.

Factor 1 = Traits associated with starch hydrolysis.

Factor 2 = Traits associated with protein modification.

Factor 3 = Traits associated with kernel plumpness.

Factor 4 = Traits associated with malt extract.

Precautions or Considerations When Using Factor Analysis

- The original sample should reflect the target population.

- Outliers should be screened for and linearity among the variables should be checked.

- Do not necessarily choose the default number of factors provided by the statistics program. Make sure that they make sense to you as classifying factors.

- Factor Analysis is an exploratory tool available to the researcher. Be careful that you don't use Factor Analysis in place of sound theoretical arguments. The Factor Analysis should be used to corroborate theory, not replace it.