DISCOVERING PATTERNS

Module 6.2 Principal Component Analysis (PCA) & Factor Analysis – Dimension Reduction

Rotated Factor Matrix[a]

	Factor				
	1	2	3	4	5
y1	.521	.059	.192	.129	.077
y2	.631	.175	.055	.114	-.079
y3	.571	.109	.119	.132	-.078
y4	.437	.129	.173	.202	.143
y5	.108	.183	.067	.527	-.017
y6	.167	.091	.160	.560	-.147
y7	.167	.053	.131	.501	-.014
y8	.163	.091	.534	.084	.005
y9	.102	.212	.608	.123	-.014
y10	.185	.155	.491	.206	.001
y11	.107	.556	.104	.155	.039
y12	.191	.552	.148	.015	-.123
y13	.094	.565	.175	.172	.104
y14	.004	.007	-.047	-.002	.204
y15	-.002	-.002	.058	-.046	.232

Extraction Method: Maximum Likelihood.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

Dr Rita Chakravarti
Institute of Systems Science
National University of Singapore
Email: rita@nus.edu.sg

Agenda

At the end of this module, you will be able to:

- Understand what is Principal Component Analysis
- Understand what is Factor Analysis
- What is Variable Clustering
- Understand how to perform these techniques using JMP, SPSS & R
- Understand why dimension reduction techniques are important for data analysis or finding patterns

PRINCIPAL COMPONENT ANALYSIS

Applicable Scenarios

Scenario 1 : A financial Analyst is interested in determining financial health of firms in a given industry – through research he has gathered about 120 relevant financial ratios. His job is to make some sense out of these 120 types of information and come up with 3-4 main indices which incorporates all those 120 ratios

Scenario 2 : A manufacturing organization collects various information of their process and they need to ensure that their processes are in control. To monitor this the quality control unit is interested in developing few key composite indices which incorporates all of those process information

Scenario 3 : A marketing manager is interested in developing a regression model to forecast sales. He has collected many information from his company which indirectly influences sales, but all of them are highly correlated (multi collinearity exists) and in turn if used in raw form will produce unstable model. He wants to create uncorrelated new variables which are linear combination of the original variables and use them to build stable model

Technique To Solve Such Problems

Principal components analysis is the technique which is used to solve such problems

It is a technique for forming new variables which are linear combination of the original variables

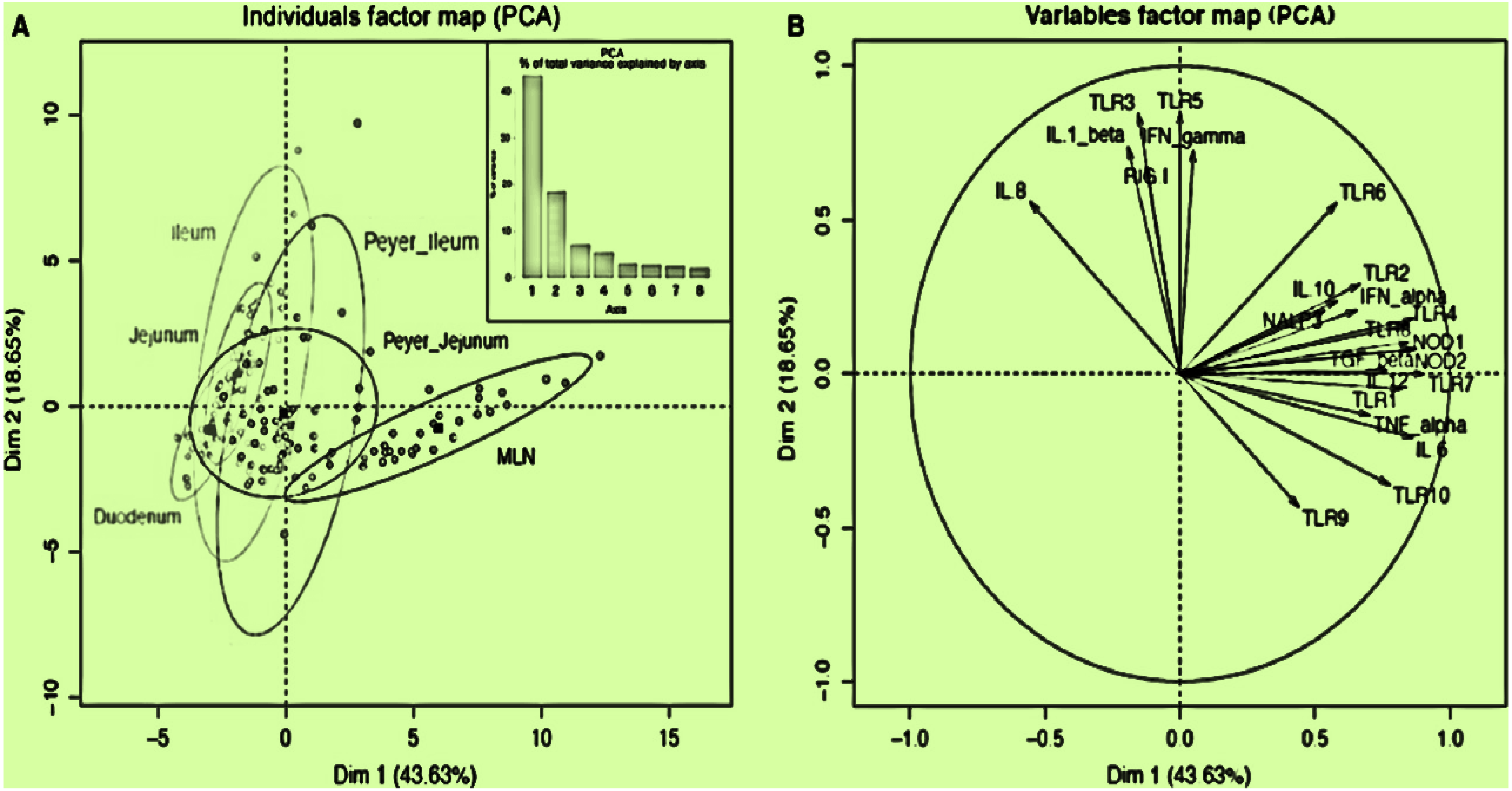
Maximum number of new variables can be created = Total number of original variables

New variables are uncorrelated among themselves

It is a dimension reduction technique

It helps identify a new set of orthogonal axes from the original set of axes

Technique To Solve Such Problems



Primary Objective of Principal Component Analysis

Geometrically the objective is to identify a new set of orthogonal axes such that :

- The coordinates of the observations with respect to each of the axes give the values of the new variables
- New axes or the variables are called principal components and the values of the new variables are called principal components score
- Each **new variable** is a **linear combination** of **original variables**
- The 1st new variable accounts for the maximum variance in the data
- The 2nd new variable accounts for the maximum variance that has not been accounted by the 1st variable
- The 3rd new variable accounts for the maximum variance that has not been accounted by the 1st two variables ……………………………………
- The pth new variable accounts for the maximum variance that has not been accounted by the p-1 variables
- The p (original no. of variables) new variables are uncorrelated

If a substantial amount of the total variance in the data is accounted for by a few variables (preferably far fewer) principal components or new variables then one can use these fewer no. of variables for further investigation instead of original p variables. This subset of variables can be identified by looking at the eigen values.

Example

Objective : Construct a measure of CPI (Consumer Price Index) for food items

Food Price Data					
City	Bread	Burger	Milk	Oranges	Tomato
Atlanta	24.5	94.5	73.9	80.1	41.6
Baltimore	26.5	91	67.5	74.6	53.3
Boston	29.7	100.8	61.4	104	59.6
Buffalo	22.8	86.6	65.3	118.4	51.2
Chicago	26.7	86.7	62.7	105.9	51.2
Cincinnati	25.3	102.5	63.3	99.3	45.6
Cleveland	22.8	88.8	52.4	110.9	46.8
Dallas	23.3	85.5	62.5	117.9	41.8
Detroit	24.1	93.7	51.5	109.7	52.4
Honolulu	29.3	105.9	80.2	133.2	61.7
Houston	22.3	83.6	67.8	108.6	42.4
Kansas City	26.1	88.9	65.4	100.9	43.2
Los Angeles	26.9	89.3	56.2	82.7	38.4
Milwakee	20.3	89.6	53.8	111.8	53.9
Minneapolis	24.6	92.2	51.9	106	50.7
New York	30.8	110.7	66	107.3	62.6
Philadelphia	24.5	92.3	66.7	98	61.7
Pittsburgh	26.2	95.4	60.2	117.1	49.3
St. Louis	26.5	92.4	60.8	115.1	46.2
San Diego	25.5	83.7	57	92.8	35.4
San Franscisco	26.3	87.1	58.3	101.8	41.5
Seattle	22.5	77.7	62	91.1	44.9
Washington, DC	24.2	93.8	66	81.6	46.2
Source : Estimated Retail Food Prices by Cities, March 1973,					
U.S. Department of Labor, Bureau of Labor Statistics,pp. 1-8.					

Example

Principal components analysis can be done either on mean corrected (X -mean) data or standardized ({X- mean}/sd) data – JMP gives standardized output

Correlations

	Bread	Burger	Milk	Oranges	Tomato
Bread	1.0000	0.6817	0.3282	0.0367	0.3822
Burger	0.6817	1.0000	0.3334	0.2109	0.6319
Milk	0.3282	0.3334	1.0000	-0.0028	0.2544
Oranges	0.0367	0.2109	-0.0028	1.0000	0.3581
Tomato	0.3822	0.6319	0.2544	0.3581	1.0000

Eigenvalues

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	2.4225	48.449		48.449
2	1.1047	22.093		70.543
3	0.7385	14.770		85.312
4	0.4936	9.872		95.185
5	0.2408	4.815		100.000

Loading Matrix

	Prin1	Prin2	Prin3	Prin4	Prin5
Bread	0.77222	-0.32437	-0.33205	0.35782	0.24529
Burger	0.89604	-0.04604	-0.22556	-0.01977	-0.37912
Milk	0.52852	-0.45280	0.71725	0.03450	-0.00387
Oranges	0.35018	0.83744	0.25059	0.33654	0.00293
Tomato	0.78823	0.30168	-0.01054	-0.50073	0.19195

Formatted Loading Matrix

	Prin1	Prin2	Prin3	Prin4	Prin5
Burger	0.896039	-0.046037	-0.225555	-0.019768	-0.379115
Tomato	0.788228	0.301677	-0.010541	-0.500729	0.191954
Bread	0.772220	-0.324370	-0.332047	0.357825	0.245290
Oranges	0.350180	0.837441	0.250592	0.336544	0.002928
Milk	0.528516	-0.452795	0.717247	0.034496	-0.003868

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5
Bread	0.49615	-0.30862	-0.38639	0.50930	0.49990
Burger	0.57570	-0.04380	-0.26247	-0.02814	-0.77264
Milk	0.33957	-0.43081	0.83464	0.04910	-0.00788
Oranges	0.22499	0.79678	0.29161	0.47902	0.00597
Tomato	0.50643	0.28703	-0.01227	-0.71271	0.39120

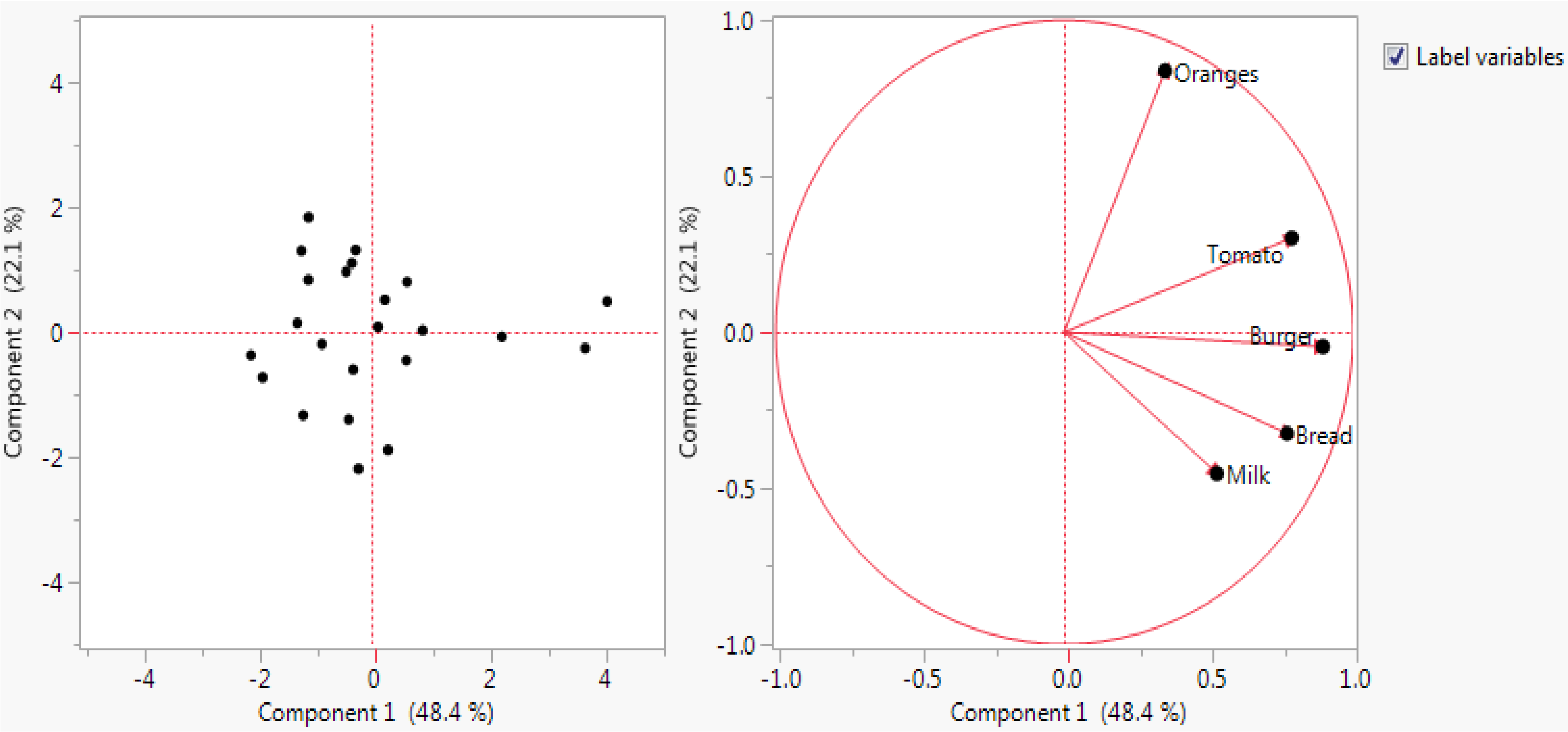
Prin1 = 0.49615 * Bread + 0.5757*Burger + 0.33957*Milk + 0.22499*Oranges + 0.50643*Tomato

Rotated Factor Loading

	Factor 1	Factor 2
Bread	0.7423941	0.0987960
Burger	0.7622247	0.3921610
Milk	0.4231628	0.0442255
Oranges	0.0228067	0.4787532
Tomato	0.4614444	0.5808128

Principal Component Output

Pictorial Representation



Appropriateness of Principal Components Analysis

Primarily depends on the objective of the study

Objective 1 : Form uncorrelated variables – but if the new variables are not interpretable then this method is not appropriate

Objective 2 : Reduce number of variables but they are linear combination of original variables – Method can only be used if there is not substantial loss of information and this depends on the purpose

Example 1 : Launch decision on space shuttle with 100 pieces of information – 99% variation is explained by 5 components but still 1% remaining variation is important

Example 2 : Consumer Price Index calculation with 100 food item prices – 99% variation is explained by 5 components and that is good enough

What is Principal Component Analysis (PCA) - Mathematically?

PCA is a robust technique which is able to produce components that are orthogonal to each other

The goal of PCA is to extract the smallest number of components which account for as much as possible of the information of the original fields.

The derived Principal Components are typically produced by forming linear combinations of the inputs, as shown by the following equations, where F_i denotes the input n fields used for the construction of the **selected** m Principal components.

$$\text{Prin 1} = a_{11} * F_1 + a_{12} * F_2 + \ldots\ldots + a_{1n} * F_n$$
$$\text{Prin 2} = a_{21} * F_1 + a_{22} * F_2 + \ldots\ldots + a_{2n} * F_n$$

$$\vdots \qquad\qquad \vdots$$

$$\text{Prin m} = a_{m1} * F_1 + a_{m2} * F_2 + \ldots + a_{mn} * F_n$$

What is Principal Component Analysis (PCA)?

The coefficients are automatically calculated by the algorithm so that the loss of information is minimal.

Components are extracted in decreasing order of importance, with the first one being the most significant as it accounts for the largest amount of the total original information.

Specifically the first component is the linear combination that carries as much as possible of the total variability of the input fields. Thus, it explains most of their information.

The second component accounts for the largest amount of the unexplained variability and is also uncorrelated with the first component. Subsequent components are constructed to account for the remaining information.

Phone Data Principal Component Analysis

INPUT FIELDS					Algorithm Generated Fields	
Customer ID	Monthly Avg No. of SMS Calls	Monthly Avg No. of MMS Calls	Monthly Avg No. of Voice Calls	Monthly Avg No. of Voice Call Minutes	Prin 1 SMS/MMS Usage	Prin 2 Voice Usage
1	19	4	90	150	-2.45	1.61
2	43	12	30	35	1.09	0.04
3	13	3	10	20	-0.19	-1.81
4	60	14	100	80	0.37	2.46
5	5	1	30	55	-1.26	-1.27
6	56	11	25	35	1.36	0.10
7	25	7	30	28	0.20	-0.78
8	3	1	65	82	-2.04	-0.28
9	40	9	15	30	0.93	-0.61
10	65	15	20	40	2.00	0.55

JMP Output for Phone Data

Principal components analysis standardized ({X- mean}/sd) data – JMP standardized output

Correlations

	Monthly Avg No. of SMS Calls	Monthly Avg No. of MMS Calls	Monthly Avg No. of Voice Calls	Monthly Avg No. of Voice Call Minutes
Monthly Avg No. of SMS Calls	1.0000	0.9807	-0.0194	-0.2281
Monthly Avg No. of MMS Calls	0.9807	1.0000	-0.0202	-0.2596
Monthly Avg No. of Voice Calls	-0.0194	-0.0202	1.0000	0.8542
Monthly Avg No. of Voice Call Minutes	-0.2281	-0.2596	0.8542	1.0000

Eigenvalues

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	2.2039	55.098		55.098
2	1.6586	41.466		96.564
3	0.1207	3.016		99.581
4	0.0168	0.419		100.000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4
Monthly Avg No. of SMS Calls	0.55886	0.42479	0.19267	-0.68564
Monthly Avg No. of MMS Calls	0.56599	0.41482	-0.02219	0.71210
Monthly Avg No. of Voice Calls	-0.36290	0.62771	-0.68161	-0.09845
Monthly Avg No. of Voice Call Minutes	-0.48541	0.50346	0.70554	0.11452

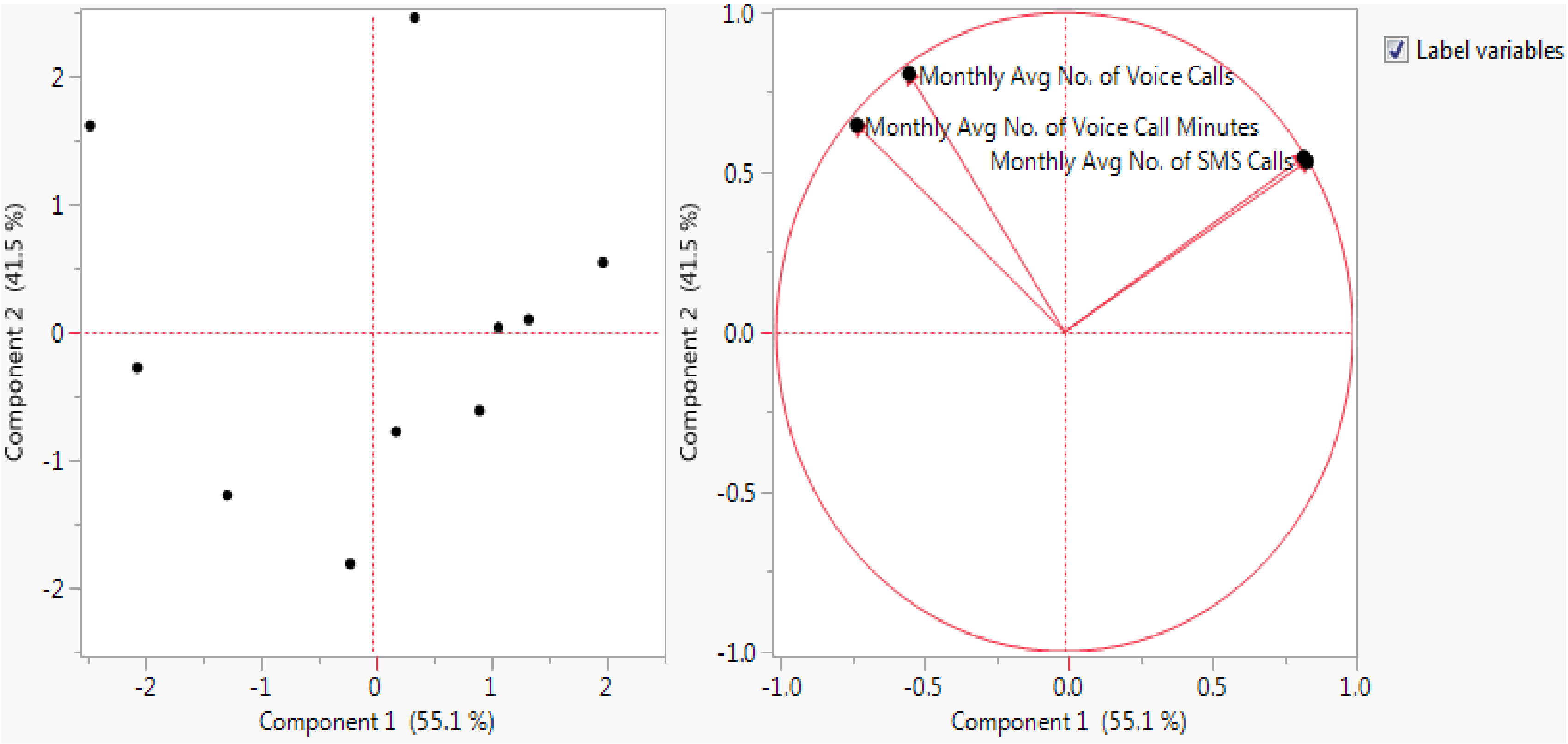
Rotated Factor Loading

	Factor 1	Factor 2
Monthly Avg No. of SMS Calls	0.983633	-0.060445
Monthly Avg No. of MMS Calls	0.985260	-0.077780
Monthly Avg No. of Voice Calls	0.039172	0.908922
Monthly Avg No. of Voice Call Minutes	-0.179656	0.903144

Prin1 = 0.5589 * Mnth avg sms + 0.5660*Mnth avg mms - 0.3629* Mnth avg no. voic – 0.4854*Mnth avg min. voic

Principal Component Output

Pictorial Representation



Principal Component Analysis (PCA)

The PCA model analyses the associations among the original fields & the identified 2 components.

Example, the SMS and MMS usage appear to be correlated and a new component was extracted to represent usage of those services.

Similarly, the number of minutes of voice calls were also correlated. The second component represents these two fields and measures the voice usage intensity.

Each derived component is standardized, with an overall population mean of zero and a standard deviation of one.

The component scores denote how many standard deviations above or below the overall mean each record stands.

In simple terms, a positive score in component 1 indicates high SMS & MMS usage while a negative score indicates below average usage.

The generated scores can then be used in subsequent modelling tasks.

KEY ISSUES in PCA Include….

How many components are to be extracted?

Is the derived solution efficient and useful?

Which original fields are mostly related with each component?

What does each component represent? What is the meaning of each component?

On How Many Components Should We Focus(extract)?

The final solution should take into account criteria such as the interpretability and business meaning of the components.

The final solution should balance simplicity with effectiveness, consisting of a reduced and interpretable set of components that can adequately represent the original fields.

Factor Analysis is similar to PCA and tends to produce results comparable to PCA.

Factor Analysis is mainly used when the main scope of the analysis is to uncover and interpret latent data dimensions whereas PCA is typically the preferred option for reducing the dimensionality of the data.

How Many Components Should Be Extracted?

- PCA Example:

Field Name	Description
Voice Out Calls	Monthly average of outgoing voice calls
Voice Out Mins	Monthly average no. of mins of out going voice calls
SMS Out Calls	Monthly average of outgoing SMS calls
MMS Out Calls	Monthly average of outgoing MMS calls
Out Calls Roaming	Monthly average of outgoing roaming calls
GPRS Traffic	Monthly average of GPRS traffic
PRC Voice Out Calls	Percentage of outgoing voice calls (outgoing voice calls as a percentage of total outgoing calls)
PRC SMS Out Calls	Percentage of outgoing SMS calls
PRC MMS Out Calls	Percentage of outgoing MMS calls
PRC Internet Calls	Percentage of outgoing internet calls
PRC Out Calls Roaming	Percentage of outgoing roaming calls (roaming calls as a percentage of total outgoing calls)

How Many Components Should Be Extracted?

- PCA applied to the telephone data reveled 5 components by using the eigenvalue criterion.

Field Name	Prin 1	Prin 2	Prin 3	Prin 4	Prin 5
Voice Out Calls	0.89	-0.34	-0.17	-0.06	-0.10
Voice Out Mins	-0.88	0.36	0.11	0.15	0.11
SMS Out Calls	0.86	-0.01	-0.16	-0.16	-0.09
MMS Out Calls	0.20	0.88	-0.04	-0.28	-0.12
Out Calls Roaming	0.26	0.86	-0.02	-0.29	-0.11
GPRS Traffic	0.19	0.09	0.60	0.35	-0.48
PRC Voice Out Calls	0.12	0.02	0.58	0.40	-0.51
PRC SMS Out Calls	0.14	0.18	-0.44	0.77	0.11
PRC MMS Out Calls	0.26	0.34	-0.46	0.66	0.08
PRC Internet Calls	0.28	0.04	0.59	0.19	0.60
PRC Out Calls Roaming	0.47	0.19	0.49	0.04	0.56

How Many Components Should Be Extracted?

Table: Eigen values & percentage of variance/information attributable to each component.

Components	Eigenvalue	Variance	Cumulative %
1	2.84	25.84	25.84
2	1.96	17.78	43.62
3	1.76	16.01	59.63
4	1.56	14.21	73.84
5	1.25	11.33	85.16
6	0.49	4.45	89.62
7	0.38	3.41	93.03
8	0.34	3.06	96.09
9	0.26	2.38	98.47
10	0.16	1.44	99.92
11	0.01	0.08	100.0

How Many Components Should Be Extracted?

The highlighted first 5 rows of the table correspond to the extracted components.

A total of 11 components are needed to fully account for the information of the 11 original fields.

Using the eigenvalue criterion, each single field contains one unit of standardized variance, components with eigenvalues below 1 are not extracted.

The second column denotes the proportion of variance attributable to each component, and the next column denotes the proportion of the variance that jointly explained by all components up to that point.

The percentage of the initial variance attributable to the 5 extracted components is 85%

How Many Components Should Be Extracted?

The eigenvalue (or latent root) criterion: Typically the eigenvalue is compared to 1 and only components with eigenvalues higher than 1 are retained.

The percentage of variance criterion: The threshold value for extraction, in general, a solution should not fall below (60%-65%) depending on the specific situation.

The interpretability and business meaning of the components: The derived components should be directly interpretable, understandable, and useful.

The scree test criterion: We should look for a large drop, followed by a "plateau" in the eigenvalue, which indicates a transition from large to small values. In other words, the first "bend" or "elbow" that indicates the maximum number of components to be extracted while the point before the bend could be selected for a more compact solution.

What is the Meaning of Each Component?

The goal is to understand the information that the component conveys and name them accordingly.

The interpretation is based on the correlations among the derived components and their original inputs by examination of the corresponding loadings.

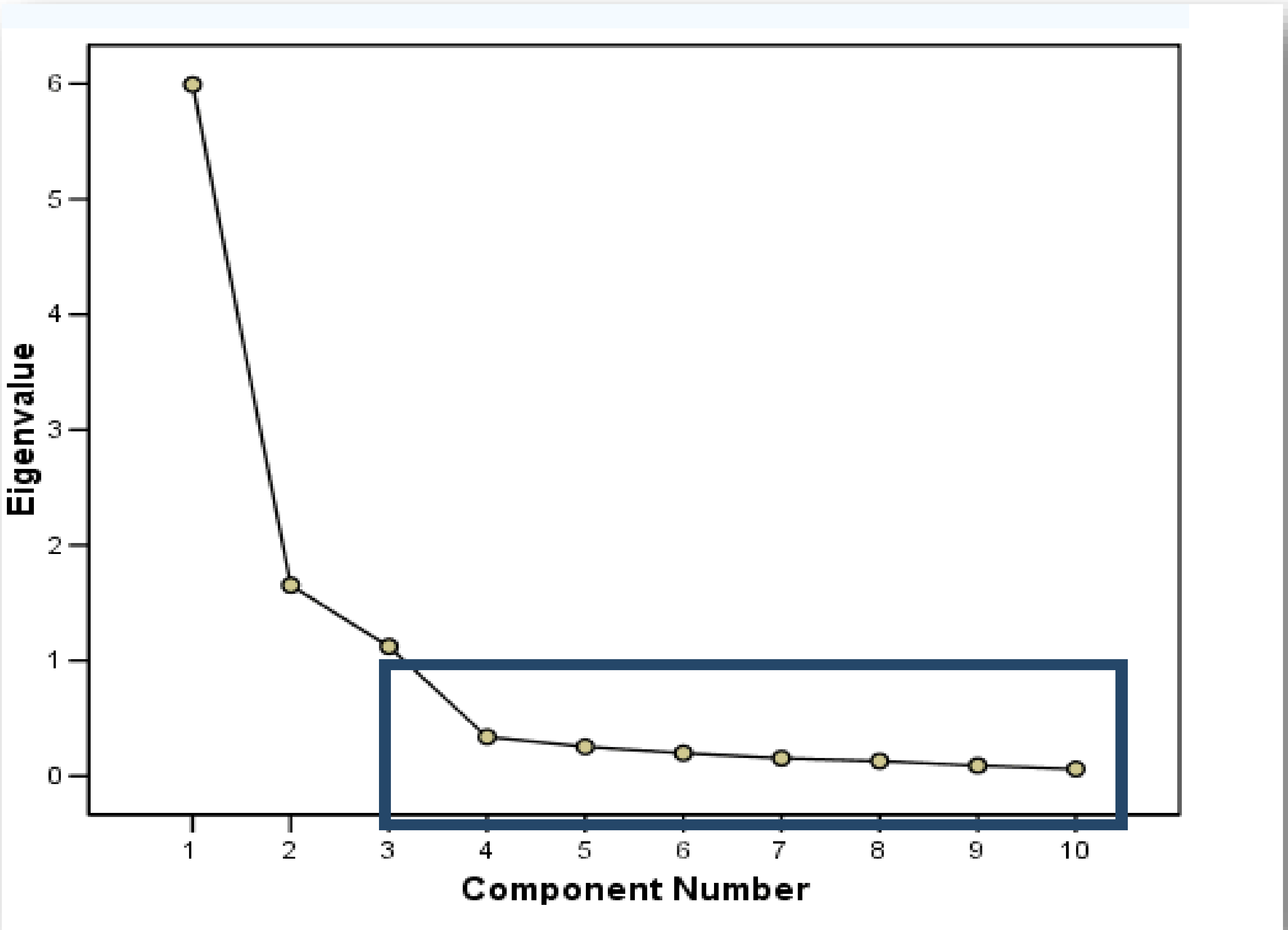
Rotation minimizes the number of fields that are strongly correlated with many components and attempts to associate each input with one component.

There are numerous rotation techniques, with the Varimax being the most popular for data reduction purposes and since it yields transparent components which are also uncorrelated.

Typically, loadings with absolute values below 0.4 are **suppressed** for easier interpretation.

Also, the original inputs are sorted according to their loadings so that fields associated with the same component appear together as a set.

Scree Plot – Helps to Decide How Many Components to Retain



What is the Meaning of Each Component?

The interpretation process involves the examination of the loading values and their signs and identification of significant correlations.

Typically correlations above 0.4 (**sometimes subjective)** in absolute value are considered to be a practical significance and denote the original fields which are representative of each component

The interpretation process **ends** with the labelling of the derived components with names that appropriately summarize their meaning.

How Many Components Should Be Extracted?

Table : Rotated Component Matrix:

Field Name	Prin 1	Prin 2	Prin 3	Prin 4	Prin 5
PRC Voice Out Calls	-0.97				
PRC SMS Out Calls	0.968				
SMS Out Calls	0.85				
Voice Out Calls		0.954			
Voice Out Mins		0.946			
PRC Out Calls Roaming			0.918		
Out Calls Roaming			0.900		
PRC MMS Out Calls				0.897	
MMS Out Calls				0.866	
PRC Internet Calls					0.880
GPRS Traffic					0.865

How do we Interpret the Components?

Component 1 is strongly associated with SMS usage. Both the number (SMS out Calls) and the ratio of SMS calls (PRC SMS out calls) load heavily on this component.

Number and minutes of voice calls (voice out calls and voice out mins) seem to covary and are combined to form the second component which we can label as voice usage.

The component 3 measures “roaming usage”.

Component 4 “MMS usage”.

Finally, the fifth component denotes “internet usage” since it represents high positive correlations with both internet calls (PRC Internet calls) and GPRS traffic (GPRS Traffic).

How can We Use the Component Scores?

The extracted components can be used in upcoming data mining models, provided that they comprise a conceptually clear and meaningful representation of the original fields.

The PCA algorithm derives new composite fields, named component scores, that denote the values of each record in the revealed components.

The derived component scores are continuous numeric fields with standardized values, hence, they have a mean of 0 and a standard deviation of 1 and they designate the deviation from the average behavior.

A list of the 5 component scores produce by PCA for 10 customers is shown in the next table.

How can we use the Component Scores?

Customer ID	Prin 1	Prin 2	Prin 3	Prin 4	Prin 5
1	0.633	-0.182	-0.263	1.346	-0.209
2	-0.964	-0.500	8.805	-0.090	-0.036
3	-0.501	-0.381	-0.196	-0.197	-0.063
4	-0.501	1.677	-0.272	-0.305	-0.055
5	3.66	-1.041	-0.385	-0.596	-0.084
6	-0.450	0.720	0.433	-0.251	-0.056
7	1.249	-0.276	1.043	-0.384	-0.028
8	-0.695	0.192	-0.204	0.461	-0.117
9	-0.902	-0.959	0.247	2.265	-0.164
10	0.028	0.212	2.715	1.186	-0.165

How can we use the Component Scores?

The high score of customer 5 in component 1 denotes a customer with above average SMS usage.

The negative score in component 2 indicates low voice usage.

Similarly, customer2 seems to be a person who frequently uses their phone abroad (roaming usage measured by component 3).

Customer 4 seems like a typical example of a “voice only” customer.

FACTOR ANALYSIS

Applicable Scenarios

A tool to measure constructs which can not be measured directly - **Attitude, Image, Intelligence, Patriotism, Sales Aptitude …..**

Scenario 1 : The marketing manager of an apparel firm wants to determine whether or not a relationship exists between patriotism and consumer's attitudes about domestic and foreign products

Scenario 2 : The president of a Fortune 500 firm wants to measure the firm's image

Scenario 3 : A sales manager is interested in measuring the aptitude of salespersons for hiring purposes

Scenario 4 : Management of a high-tech firm is interested in measuring determinants of resistance to technological innovations

Initial Application of Factor Analysis

Originally developed to explain student performance in various courses and to understand the link between grades and intelligence

Observable measure : Student's grade in various subjects

Assumption : Student's grade in any course is a function of : 1) General Intelligence Level (I) and 2) Aptitude for that course (A)

Regression equations : M : Math Score P : Phy Score C : Chem Score E : Eng Score H : Hist Score F : French Score

A-s are the aptitude/error term of the regression equation. 0.8, 0.7 … pattern loadings

I : Latent Factor

$M = .8\,I + A_m \quad P = .7\,I + A_p$

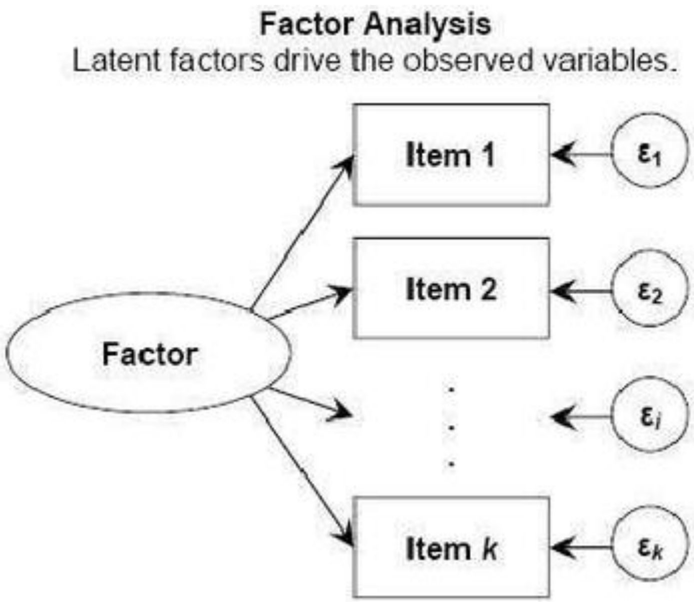
$C = .9\,I + A_c \quad E = .6\,I + A_e$

$H = .5\,I + A_h \quad F = .65\,I + A_f$

What is Factor Analysis?

Conceptual Model:

Factor: Latent variables qualities that you cannot measure directly.



Item j : Observed variables qualities that you can measure directly.

- Item i - s are only related to each other through their common relationship with Factor.
- A factor consists of relatively homogeneous variables.

What is Factor Analysis?

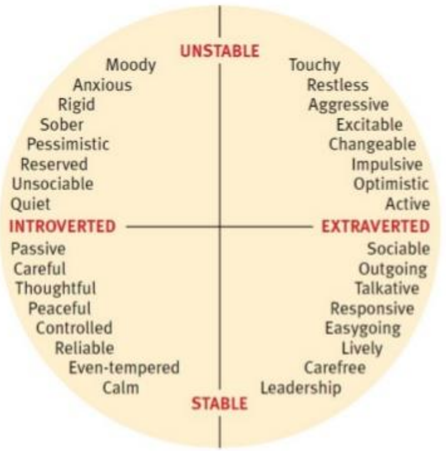
You have a very large dataset, with many, many variables and you want to discover latent factors

These variables could be highly correlated

You wish to reduce the number of variables, to assist in

- Visual analysis and exploration of the data
- To identify natural groupings or summaries of the data
- To create a set of latent factors that can later be used to create predictive models

Factor Analysis and the Eysencks' Personality Dimensions



- Factor Analysis: Identifying factors that tend to cluster together.
- Using factor analysis, Hans and Sybil Eysenck found that many personality traits actually are a function of two basic dimensions along which we all vary.
- Research supports their idea that these variations are linked to genetics.

What is Factor Analysis?

Factor Analysis aims at effectively reducing the data's dimensions and removing redundant information. They do so by replacing the initial set of inputs with a core set of compound measures (factors) which simplify subsequent modelling while retaining most of the information of the original attributes.

Factor analysis examines the correlations among attributes to identify these basic dimensions.

If certain continuous inputs/attributes tend to co-vary then they are correlated. If their relationship is expressed adequately by a straight line then they have a strong linear correlation and these attributes are mapped to representative fields named factors or components.

The derived components and factors have the form of continuous numeric scores and can be subsequently used as any other fields for reporting or modelling purposes.

The factors/components are named by examining "factor loadings" which are estimates of the correlations between attribute ratings and perception measures. The interpretation of the factors/components is an essential part of the data reduction procedure since the derived factors will be used in subsequent tasks, it is important to fully understand the information they convey.

Comment : Factor analysis can be considered as extension of principal component analysis. However the approximation based on the factor analysis model is more elaborate

Factor Analysis Techniques : Most popular : a) Principal Components Factoring (PCF) b) Principal Axis Factoring (PAF)

Factor Matrix

A factor matrix shows variables in rows and factors in columns

Columns represent derived factors

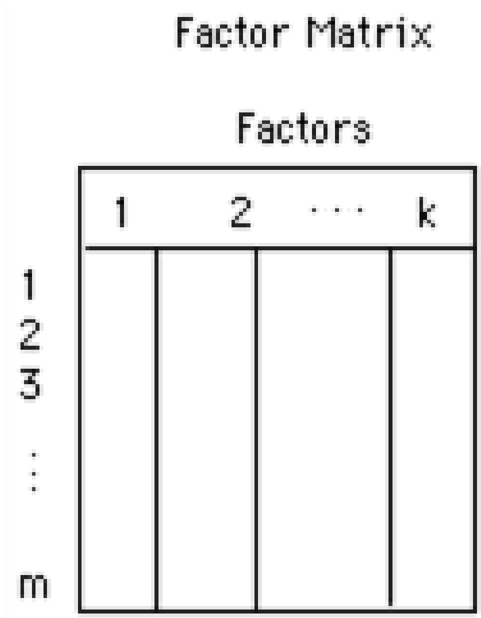
Rows represent input variables

Loadings represent degree to which each of the variables “correlates” with each of the factors

Loadings range from -1 to 1

Inspection of factor loadings reveals extent to which each of the variables contributes to the meaning of each of the factors.

High loadings provide meaning and interpretation of factors (~ regression coefficients)



What is Factor Analysis?

Imagine you have collected many , many variables regarding sales in a major department store

- Types of items purchased by individual customers
- Amount spent by individual customers
- Payment method used by individual customers
- Age of individual customers
- …………..
- You know that many of the variables are correlated, e.g. Amount spent by individual customers will be correlated with the payment method used by the customers

You would want to reduce set of variables to a smaller number of uncorrelated variables; thus allowing you to

- Make analysis and interpretation of data more easy
- See which variables and combinations of variables are related to each other

Factor Analysis is intended to do this, by transforming an original set of variables to a smaller number of uncorrelated variables which are a linear combination of the original variables – These represent the core dimensions of the consumer behavior

Benefits of Factor Analysis

Reduction of dataset to a smaller set of uncorrelated variables through the removal of irrelevant or repeated data

Some data mining techniques will run too slow or not at all if they have to handle a large number of inputs. (Situations like these can be avoided by using the derived factors instead of the many original inputs)

Simplicity is the key benefit of data reduction techniques since they drastically reduce the number of fields under study to a core set of factors.

Easier to visualize the data and relationships between the data

Benefits of Factor Analysis

Easier to fit models that relate the dataset to some other response variable

Deeper understanding of the relative importance of the variables that comprise a dataset

Factor Analysis is an unsupervised, statistical techniques that are among the most popular data reduction techniques.

Additionally, clustering solutions can also be biased if the inputs are dominated by correlated "variants" of the same attribute. By using a data reduction technique we can unveil the true data dimensions and ensure that they are equal weight in the formation of the final clusters.

Illustration - One Factor Model

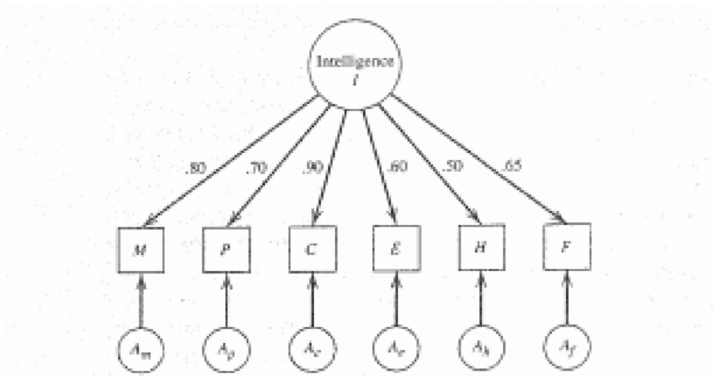
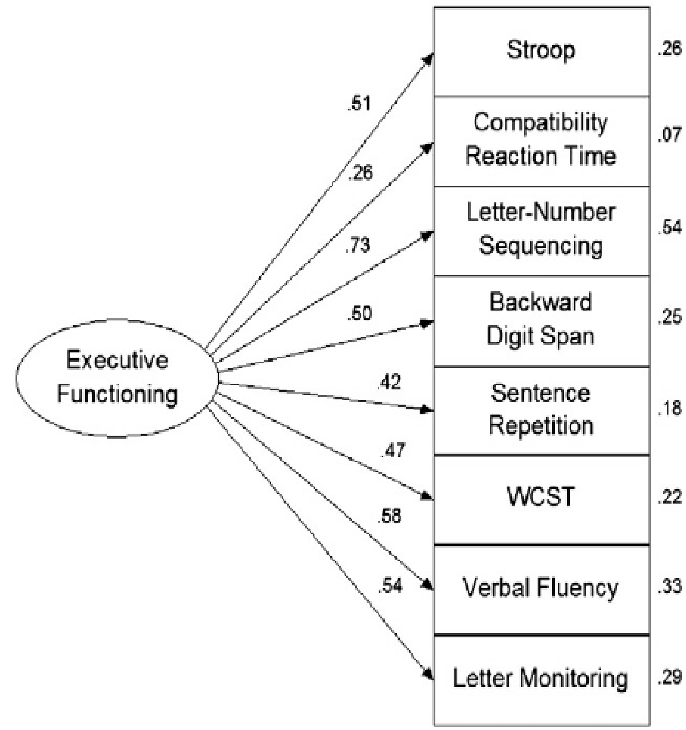
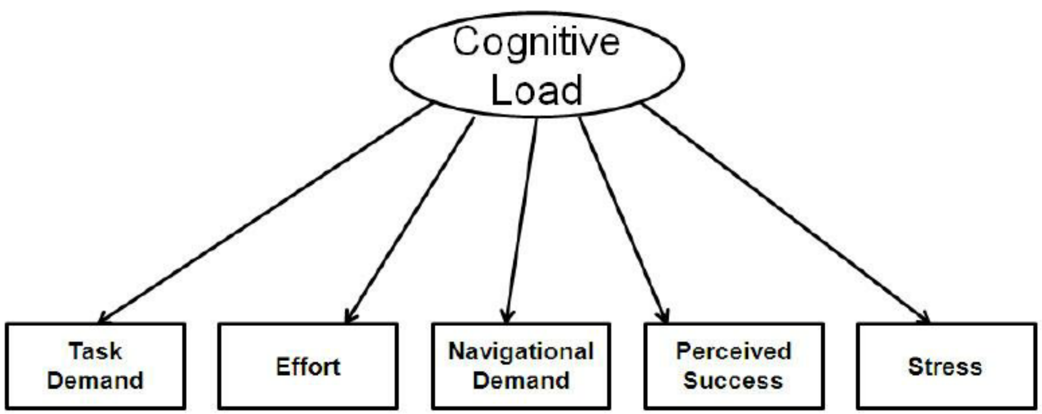


Figure 5.1 Relationship between grades and intelligence.





Two Factor Model

Extension to two factor model

Observable measure : Student's grade in various subjects

Assumption : Student's grade in any course is a function of 1) Factor 1 (Q) 2) Factor 2 (V) and 3) Aptitude for that course (A)

Regression equations : M : Math Score P : Phy Score C : Chem Score E : Eng Score H : Hist Score F : French Score

A-s are aptitude and error term of the regression equations. 0.8, 0.7 … pattern loadings,

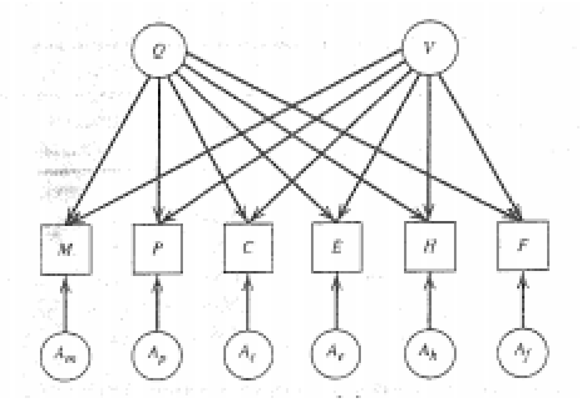
Q & V : Latent Factors Q : Quantitative Ability V: Vernacular Ability

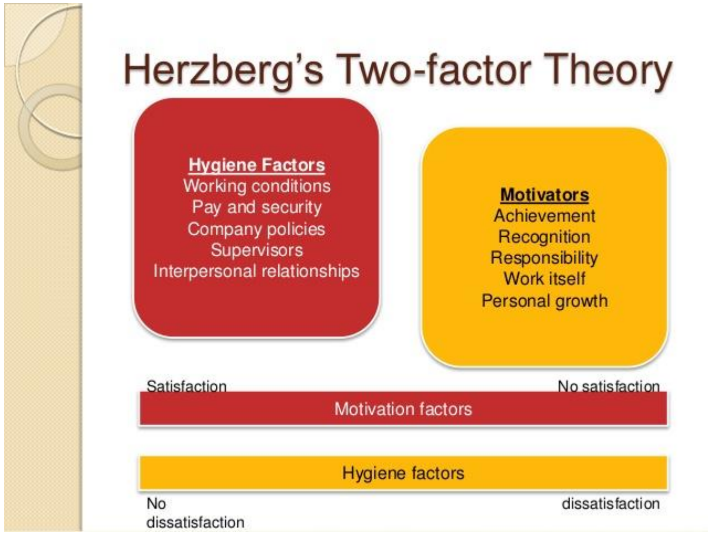
$M = .8\,Q + .2\,V + A_m$ $\quad P = .7\,Q + .3\,V + A_p$

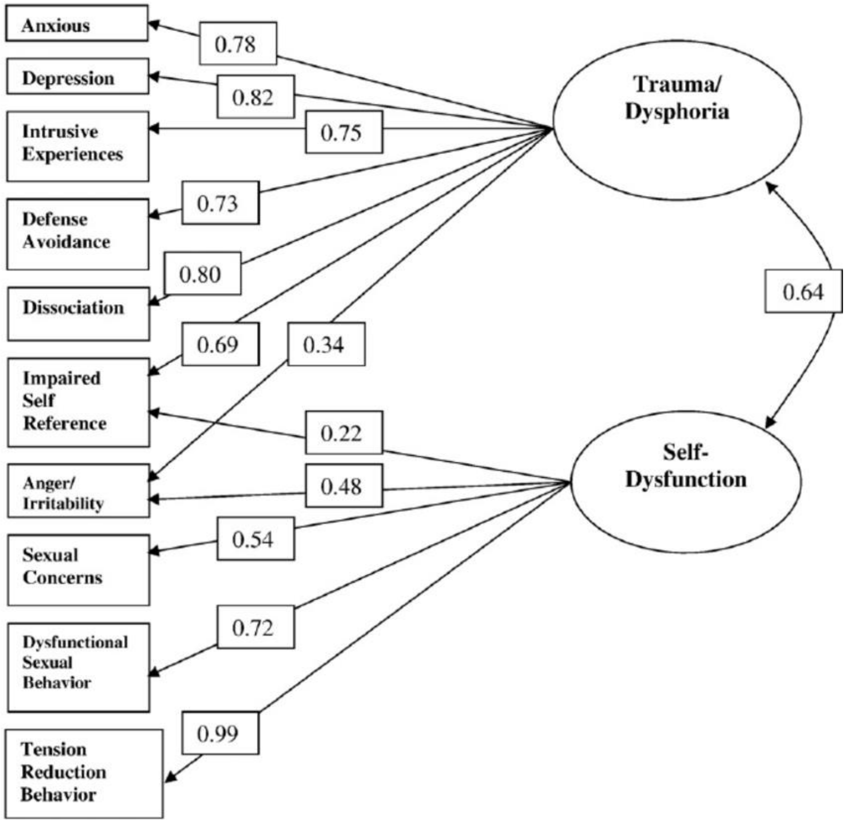
$C = .6\,Q + .3\,V + A_c$ $\quad E = .2\,Q + .8\,V + A_e$

$H = .15\,Q + .82\,V + A_h$ $\quad F = .25\,Q + .85\,V + A_f$

Illustration - Two Factor Model







Generalized Mathematical Expression of Factor Analysis

m-factor model with p indicators

$$x_1 = \lambda_{11}\,\xi_1 + \lambda_{12}\,\xi_2 + \ldots\ldots\ldots\ldots\ldots\ldots + \lambda_{1m}\,\xi_m + \varepsilon_1$$
$$x_2 = \lambda_{21}\,\xi_1 + \lambda_{22}\,\xi_2 + \ldots\ldots\ldots\ldots\ldots\ldots + \lambda_{2m}\,\xi_m + \varepsilon_2$$

.
.
.

$$x_p = \lambda_{p1}\,\xi_1 + \lambda_{p2}\,\xi_2 + \ldots\ldots\ldots\ldots\ldots\ldots + \lambda_{pm}\,\xi_m + \varepsilon_p$$

The inter- correlation among p variables are explained by m latent factors. Typically m is much less than p. If m factors are uncorrelated they are called orthogonal model. If they are correlated they are called oblique model

Factor Analysis vs Principal Components Analysis

Both are data reduction techniques

PCA	FA
PCA Objective : Reduce the number of variables to a few components such that each component forms a new variable and the number of retained components explains the maximum amount of variance in the data	FA Objective : Search or identify underlying factors(s) or latent constructs that can explain the inter-correlation among the variables
Emphasis is on explaining variance in the data	Emphasis is on explaining the correlation among the indicators
Variables form an index e.g. Consumer price index	Variables reflect presence of the unobservable construct or factor
Variables are called formative indicators	Variables are called reflective indicators

Exploratory vs Confirmatory Factor Analysis & Rotation

In exploratory factor analysis the researcher has little or no knowledge about factor structure e.g. excellence of a given firm → data exploration is necessary

In confirmatory factor analysis assumes that the factor structure is known

Rotation : The objective of rotation is to achieve a simpler factor structure that can be meaningfully interpreted as a latent factor

Different Rotation Methods :

- **Varimax** : Major objective is to have a factor structure in which each variable loads highly on one and only one factor (i.e. high loading in one factor and near Zero on other factors) - **Most popular**
- **Quartimax** : All the variables have a fairly high loading on one factor. Each Variable should have a high loading on one other factor and near zero loadings on the remaining factors
- **Equimax** : Combination of Varimax & Quartimax
- **Direct Oblimin :** Method in which factors are allowed to be correlated
- **Promax** : Similar to Direct Oblimin but faster – used for large data sets

Factor Extraction – Real Life Example of Variables – Case 1 – Total 304 Variables

Table 7: Information Value (Infoval)

	VARIABLE NAME	INFOVAL
	CUST – AVG TOTBAL[1-3] % LIMIT (CUST930A)	87.8
	AVG UTIL TOTBAL[1-3] % LIMIT (CHRVA910A)	85.3
	AVG TOTBAL[1-3] % LIMIT (CHRVA930A)	85.0
	CUST – AVG TOTBAL[1-6] % LIMIT (CUST931A)	83.2
	TOTBAL[1] % LIMIT (CHRVA350A)	82.8
	AVG UTIL TOTBAL[1-6] % LIMIT (CHRVA911A)	81.6
	AVG TOTBAL[1-6] % LIMIT (CHRVA931A)	80.9
	CUST – AVG BAL[1-3] % LIMIT (CUST930)	79.1
	BAL[1] % LIMIT (CHRVA350)	77.1
0	CUST – AVG TOTBAL[1-12] % LIMIT (CUST932A)	75.7
1	AVG UTIL TOTBAL[1-12] % LIMIT (CHRVA912A)	75.4
2	AVG UTIL BAL[1-3] % LIMIT (CHRVA910)	75.3
3	AVG BAL[1-3] % LIMIT (CHRVA930)	75.1
4	AVG TOTBAL[1-12] % LIMIT (CHRVA932A)	73.7
5	CUST – AVG BAL[1-6] % LIMIT (CUST931)	73.5
6	AVG UTIL BAL[1-6] % LIMIT (CHRVA911)	70.7
7	AVG BAL[1-6] % LIMIT (CHRVA931)	69.9
8	CUST – TOTBAL[2-6] % PAY[1-5] (CUST412B)	69.0
9	CUST – PAY[1-5] % BAL[2-6] (CUST412)	67.2
0	CUST – AVG BAL[1-12] % LIMIT (CUST932)	66.3
1	CUST – PAY[1-5] % TOTBAL[2-6] (CUST412A)	65.4
2	CUST – PAY[1-11] % BAL[2-12] (CUST413)	65.3
3	CUST – PAY[1-11] % TOTBAL[2-12] (CUST413A)	65.2
4	AVG UTIL BAL[1-12] % LIMIT (CHRVA912)	64.5
5	AVG BAL[1-12] % LIMIT (CHRVA932)	63.3
	VARIABLE NAME	**INFOVAL**
6	CUST – PAY[1-3] % BAL[2-4] (CUST411)	61.9
7	PAY[1-5] % TOTBAL[2-6] (CHRVA412A)	61.2
8	PAY[1-11] % BAL[2-12] (CHRVA413)	60.7
9	PAY[1-5] % BAL[2-6] (CHRVA412)	60.7
0	PAY[1-11] % TOTBAL[2-12] (CHRVA413A)	60.7
1	PAY[1-5] % AMTDUE[2-6] (CHRVA422)	60.2
2	CUST – PAY[1-3] % TOTBAL[2-4] (CUST411A)	59.6
3	CUST – PURCHASE AMT[1-6] % BAL[1-6] (CUST462)	59.4
4	CUST – PURCHASE AMT[1-3] % BAL[1-3] (CUST461)	58.7
5	PAY[1-3] % AMTDUE[2-4] (CHRVA421)	58.6
6	CUST – PURCHASE AMT[1-12] % BAL[1-12] (CUST463)	57.5
7	PURCHASE AMT[1-6] % BAL[1-6] (CHRVA462)	57.3
8	CUST – PURCHASE AMT[1-6] % TOTBAL[1-6] (CUST462A)	56.9
9	PURCHASE AMT[1-3] % BAL[1-3] (CHRVA461)	56.3
0	PURCHASE AMT[1-6] % TOTBAL[1-6] (CHRVA462A)	56.1
1	PAY[1-11] % AMTDUE[2-12] (CHRVA423)	55.8
2	CUST – PURCHASE AMT[1-12] % TOTBAL[1-12] (CUST463A)	55.7
3	PAY[1-3] % TOTBAL[2-4] (CHRVA411A)	55.7
4	CUST – PURCHASE AMT[1-3] % TOTBAL[1-3] (CUST461A)	55.5
5	PURCHASE AMT[1-12] % BAL[1-12] (CHRVA463)	55.4
6	PAY[1-3] % BAL[2-4] (CHRVA411)	54.4
7	PURCHASE AMT[1-12] % TOTBAL[1-12] (CHRVA463A)	54.2
8	BAL[1] % MAX BAL LF (CHRVA346)	53.4
9	PURCHASE AMT[1-3] % TOTBAL[1-3] (CHRVA461A)	53.3
0	CASH BAL[1] % AVG CASH BAL[1-6] (CHRVA490)	51.8
1	AVG BAL[1-3] (CHRVA310)	51.4
2	AVG TOTBAL[1-3] (CHRVA310A)	49.7
3	AVG BAL[1-6] (CHRVA311)	49.2
4	BAL[1] (CHRVA300)	49.0
5	AVG TOTBAL[1-6] (CHRVA311A)	48.5
6	TOT CASH BAL[1] % AVG TOT CASH BAL[1-6] (CHRVA490A)	48.4
7	AVG TOTBAL[1-12] (CHRVA312A)	47.9
8	CASH BAL[1] % LIMIT (MISS003)	47.5
9	PCNT (TOTBAL > LIMIT) [1-12] (CHRVA937A)	46.3

Factor Extraction – Real Life Example of Variables – Case 1 - continued

0	AVG BAL[1-12] (CHRVA312)	46.3
1	TOTBAL[1] % AVG TOTBAL[1-6] (CHRVA341A)	46.2
2	MAX BAL[1-3] (CHRVA320)	46.1
3	TOTBAL[1] % AVG TOTBAL[2-6] (CHRVA904A)	46.1
4	TIMES TOT OVL[1-12] (MISS011A)	45.8
5	BAL[1] % BAL[2] (CHRVA340)	45.7
6	TOTBAL[1] (CHRVA300A)	45.7
7	TOT CASH BAL[1] % LIMIT (MISS003A)	45.4
8	CUST – TOTBAL[1] % AVG TOTBAL[1-6] (CUST341A)	45.2
9	PCNT (TOTBAL > LIMIT) [1-6] (CHRVA936A)	44.5
0	TIMES TOT OVL[1-6] (MISS010A)	44.3
1	MAX TOTBAL[1-3] (CHRVA320A)	43.9
2	MOS SINCE FULL PAY[1-12] (CHRVA555)	43.7
3	CASH BAL[1] % TOTBAL[1] (CHRVA480A)	43.3
4	CASH BAL[1] % BAL[1] (CHRVA480)	43.3
5	CUST – BAL[1] % AVG BAL[1-6] (CUST341)	42.7
6	CASH BAL[1-3] % BAL[1-3] (CHRVA481)	42.7
7	TOTBAL[1] % TOTBAL[2] (CHRVA340A)	42.7
8	CASH BAL[1-3] % TOTBAL[1-3] (CHRVA481A)	42.6
9	TOTBAL[1] % AVG TOTBAL[2-12] (CHRVA905A)	42.5
0	TOTBAL[1] % AVG TOTBAL[1-12] (CHRVA342A)	42.2
1	CUST – PURCHASE AMT[1] % BAL[1] (CUST460)	41.8
2	CUST – TOTBAL[1] % AVG TOTBAL[1-12] (CUST342A)	41.8
3	BAL[1] % AVG BAL[2-6] (CHRVA904)	41.7
4	PCNT (TOTBAL > LIMIT) [1-3] (CHRVA935A)	41.5
5	BAL[1] % AVG BAL[1-6] (CHRVA341)	41.4
6	CASH BAL[1-6] % BAL[1-6] (CHRVA482)	41.4
7	CASH BAL[1-6] % TOTBAL[1-6] (CHRVA482A)	41.3
	VARIABLE NAME	**INFOVAL**
8	TIMES TOT OVL[1-3] (MISS009A)	41.2
9	MAX BAL[1-6] (CHRVA321)	40.4
0	CASH BAL[1-12] % BAL[1-12] (CHRVA483)	40.2
1	TOT CASH BAL[1-3] % TOTBAL[1-3] (CHRVA481B)	40.1
2	CASH BAL[1-12] % TOTBAL[1-12] (CHRVA483A)	40.0
3	CUST – PURCHASE AMT[1] % TOTBAL[1] (CUST460A)	39.5
4	TOT CASH BAL[1] % TOTBAL[1] (CHRVA480B)	39.4
5	TOT CASH BAL[1-6] % TOTBAL[1-6] (CHRVA482B)	39.3
6	TOT CASH BAL[1-12] % TOTBAL[1-12] (CHRVA483B)	38.9
7	BAL[1] % MAX BAL[1-12] (CHRVA345)	38.6
8	CUST – BAL[1] % AVG BAL[1-12] (CUST342)	38.6
9	PURCHASE AMT[1] % BAL[1] (CHRVA460)	38.4
00	BAL[1] % AVG BAL[2-12] (CHRVA905)	38.4
01	TOTBAL[1] % MAX TOTBAL[1-12] (CHRVA345A)	38.4
02	PCNT (BAL > LIMIT) [1-12] (CHRVA937)	38.2
03	BAL[1] % AVG BAL[1-12] (CHRVA342)	38.1
04	TIMES OVL[1-12] (MISS011)	37.9
05	TIMES OVL[1-6] (MISS010)	37.8
06	PCNT (BAL > LIMIT) [1-6] (CHRVA936)	37.8
07	MAX TOTBAL[1-6] (CHRVA321A)	37.5
08	PCNT (BAL > LIMIT) [1-3] (CHRVA935)	36.0
09	TIMES OVL[1-3] (MISS009)	35.8
10	PURCHASE AMT[1] % TOTBAL[1] (CHRVA460A)	35.6
11	PCNT PAID IN FULL OVER 23 MOS – DLQ_HIST[1-23] = B (MISS001)	35.5
12	MAX BAL[1-12] (CHRVA322)	34.3
13	MAX TOTBAL[1-12] (CHRVA322A)	31.5
14	CASH AMT[1-12] % PURCHASE AMT[1-12] (CHRVA473)	29.3
15	BAL[1] % MAX BAL[1-6] (CHRVA344)	29.1
16	TOTBAL[1] % MAX TOTBAL[1-6] (CHRVA344A)	28.7
17	CASH AMT[1-3] % PURCHASE AMT[1-3] (CHRVA471)	26.6
18	CASH AMT[1-6] % PURCHASE AMT[1-6] (CHRVA472)	26.5
19	PCNT CASH AMT[1-12] > 0 (CHRVA475)	23.0
20	NUM CASH ADV[1-12] (MISS008)	23.0
21	MAX BAL[1-6] (CHRVA321)	40.4
22	CASH BAL[1-12] % BAL[1-12] (CHRVA483)	40.2

Final Selection of Variables for Modeling – Case1

Out of 304 variables – Only 7 variables are used for final prediction :

- Avg payment / avg balance
- Avg cash balance / avg total balance
- % delinquency in last 12 mnths
- % worst delinquency in last 3 mnths
- Max purchase amount in last 3 mnths / Total purchase amounts in 3 mnths
- Mnths since maximum balance in last 12 mnths
- Maximum consecutive increase in total balance in last 12 mnths

Factor Analysis – Important Factors – Case 2 – Total 1000 variables

```
The FACTOR Procedure
Initial Factor Method: Principal Components

Prior Communality Estimates: ONE


Eigenvalues of the Covariance Matrix: Total = 981482.115  Average = 7549.86242
```

	Eigenvalue	Difference	Proportion	Cumulative
1	391723.451	170693.062	0.3991	0.3991
2	221030.389	38650.129	0.2252	0.6243
3	182380.261	24644.559	0.1858	0.8101
4	157735.702	129312.086	0.1607	0.9708
5	28423.616	28366.352	0.0290	0.9998
6	57.264	29.579	0.0001	0.9999
7	27.685	7.581	0.0000	0.9999
8	20.103	7.858	0.0000	0.9999
9	12.245	0.618	0.0000	0.9999
10	11.627	1.264	0.0000	0.9999
11	10.363	3.521	0.0000	0.9999
12	6.842	1.420	0.0000	1.0000
13	5.422	0.629	0.0000	1.0000

Factor Analysis – Important Factors – Case 2 – Total 1000 variables

Equivalent variables in Factor 1:

Number of Times Bucket 5+ for Non-ABC Credit Cards in Last 6 months	f_m5_6_cc	0.97356
Number of Times Bucket 6+ for Non-ABC Credit Cards in Last 6 months	f_m6_6_cc	0.97356
Number of Times Bucket 5+ for Non-ABC Credit Cards in Last 3 months	f_m5_3_cc	0.97356
Number of Times Bucket 4+ for Non-ABC Credit Cards in Last 6 months	f_m4_6_cc	0.97355
Number of Times Bucket 4+ for Non-ABC Credit Cards in Last 3 months	f_m4_3_cc	0.97355
Number of Times Bucket 6+ for Non-ABC Credit Cards in Last 3 months	f_m6_3_cc	0.97355
Number of Times Bucket 6+ for Non-ABC Credit Cards in Last 9 months	f_m6_9_cc	0.97355
Number of Times Bucket 5+ for Non-ABC Credit Cards in Last 9 months	f_m5_9_cc	0.97355
Number of Times Bucket 3+ for Non-ABC Credit Cards in Last 3 months	f_m3_3_cc	0.97354
Number of Times Bucket 6+ for Non-ABC Credit Cards in Last 12 months	f_m6_12_cc	0.97354
Number of Times Bucket 5+ for Non-ABC Credit Cards in Last 12 months	f_m5_12_cc	0.97354
Number of Times Bucket 4+ for Non-ABC Credit Cards in Last 9 months	f_m4_9_cc	0.97353
Number of Times Bucket 3+ for Non-ABC Credit Cards in Last 6 months	f_m3_6_cc	0.97353
Number of Times Bucket 4+ for Non-ABC Credit Cards in Last 12 months	f_m4_12_cc	0.97352
Number of Times Bucket 2+ for Non-ABC Credit Cards in Last 3 months	f_m2_3_cc	0.97352
Number of Times Bucket 3+ for Non-ABC Credit Cards in Last 9 months	f_m3_9_cc	0.97350
Number of Times Bucket 3+ for Non-ABC Credit Cards in Last 12 months	f_m3_12_cc	0.97348
Number of Times Bucket 2+ for Non-ABC Credit Cards in Last 6 months	f_m2_6_cc	0.97347
Latest Maximum DPD for Non-ABC Credit Cards	f_ldpd_cc	0.97346
Number of Times Bucket 1+ for Non-ABC Credit Cards in Last 3 months	f_m1_3_cc	0.97342
Maximum DPD for Non-ABC Credit Cards in Last 3 months	mxdpd_3_cc	0.97342
Number of Times Bucket 2+ for Non-ABC Credit Cards in Last 9 months	f_m2_9_cc	0.97339
Maximum DPD for Non-ABC Credit Cards in Last 6 months	mxdpd_6_cc	0.97338
Maximum DPD for Non-ABC Credit Cards in Last 9 months	mxdpd_9_cc	0.97335
Number of Times Bucket 2+ for Non-ABC Credit Cards in Last 12 months	f_m2_12_cc	0.97335
Maximum DPD for Non-ABC Credit Cards in Last 12 months	mxdpd_12_cc	0.97334
Number of Times Bucket 1+ for Non-ABC Credit Cards in Last 6 months	f_m1_6_cc	0.97327
Number of Times Bucket 1+ for Non-ABC Credit Cards in Last 9 months	f_m1_9_cc	0.97310
Number of Times Bucket 1+ for Non-ABC Credit Cards in Last 12 months	f_m1_12_cc	0.97298

Factor Analysis – Important Factors – Case 2 – Total 1000 variables

Equivalent variables in Factor 2:

Maximum DPD for Non-ABC Hire Purchases in Last 3 months	mxdpd_3_hp	0.98974
Latest Maximum DPD for Non-ABC Hire Purchases	f_ldpd_hp	0.98974
Maximum DPD for Non-ABC Hire Purchases in Last 6 months	mxdpd_6_hp	0.98973
Maximum DPD for Non-ABC Hire Purchases in Last 9 months	mxdpd_9_hp	0.98973
Maximum DPD for Non-ABC Hire Purchases in Last 12 months	mxdpd_12_hp	0.98973
Number of Times Bucket 4+ for Non-ABC Hire Purchases in Last 6 months	f_m4_6_hp	0.98973
Number of Times Bucket 3+ for Non-ABC Hire Purchases in Last 6 months	f_m3_6_hp	0.98972
Number of Times Bucket 5+ for Non-ABC Hire Purchases in Last 6 months	f_m5_6_hp	0.98972
Number of Times Bucket 3+ for Non-ABC Hire Purchases in Last 3 months	f_m3_3_hp	0.98972
Number of Times Bucket 2+ for Non-ABC Hire Purchases in Last 3 months	f_m2_3_hp	0.98972
Number of Times Bucket 6+ for Non-ABC Hire Purchases in Last 6 months	f_m6_6_hp	0.98971
Number of Times Bucket 4+ for Non-ABC Hire Purchases in Last 9 months	f_m4_9_hp	0.98971
Number of Times Bucket 5+ for Non-ABC Hire Purchases in Last 9 months	f_m5_9_hp	0.98971
Number of Times Bucket 4+ for Non-ABC Hire Purchases in Last 3 months	f_m4_3_hp	0.98970
Number of Times Bucket 6+ for Non-ABC Hire Purchases in Last 9 months	f_m6_9_hp	0.98970
Number of Times Bucket 4+ for Non-ABC Hire Purchases in Last 12 months	f_m4_12_hp	0.98970
Number of Times Bucket 5+ for Non-ABC Hire Purchases in Last 3 months	f_m5_3_hp	0.98970
Number of Times Bucket 5+ for Non-ABC Hire Purchases in Last 12 months	f_m5_12_hp	0.98970
Number of Times Bucket 6+ for Non-ABC Hire Purchases in Last 3 months	f_m6_3_hp	0.98969
Number of Times Bucket 6+ for Non-ABC Hire Purchases in Last 12 months	f_m6_12_hp	0.98969
Number of Times Bucket 3+ for Non-ABC Hire Purchases in Last 9 months	f_m3_9_hp	0.98967
Number of Times Bucket 1+ for Non-ABC Hire Purchases in Last 3 months	f_m1_3_hp	0.98967
Number of Times Bucket 3+ for Non-ABC Hire Purchases in Last 12 months	f_m3_12_hp	0.98965
Number of Times Bucket 2+ for Non-ABC Hire Purchases in Last 6 months	f_m2_6_hp	0.98964
Number of Times Bucket 2+ for Non-ABC Hire Purchases in Last 9 months	f_m2_9_hp	0.98945
Number of Times Bucket 1+ for Non-ABC Hire Purchases in Last 6 months	f_m1_6_hp	0.98945
Number of Times Bucket 2+ for Non-ABC Hire Purchases in Last 12 months	f_m2_12_hp	0.98936
Number of Times Bucket 1+ for Non-ABC Hire Purchases in Last 9 months	f_m1_9_hp	0.98905
Number of Times Bucket 1+ for Non-ABC Hire Purchases in Last 12 months	f_m1_12_hp	0.98886

Factor Analysis – Important Factors – Case 2 – Total 1000 variables

Equivalent variables in Factor 3:

Number of Times Bucket 4+ for Non-ABC Housing Loans in Last 6 months	f_m4_6_hl	0.98532
Number of Times Bucket 5+ for Non-ABC Housing Loans in Last 6 months	f_m5_6_hl	0.98531
Number of Times Bucket 3+ for Non-ABC Housing Loans in Last 6 months	f_m3_6_hl	0.98531
Maximum DPD for Non-ABC Housing Loans in Last 3 months	mxdpd_3_hl	0.98530
Number of Times Bucket 6+ for Non-ABC Housing Loans in Last 6 months	f_m6_6_hl	0.98530
Latest Maximum DPD for Non-ABC Housing Loans	f_ldpd_hl	0.98530
Maximum DPD for Non-ABC Housing Loans in Last 6 months	mxdpd_6_hl	0.98529
Number of Times Bucket 4+ for Non-ABC Housing Loans in Last 9 months	f_m4_9_hl	0.98529
Number of Times Bucket 3+ for Non-ABC Housing Loans in Last 3 months	f_m3_3_hl	0.98529
Number of Times Bucket 5+ for Non-ABC Housing Loans in Last 9 months	f_m5_9_hl	0.98529
Number of Times Bucket 2+ for Non-ABC Housing Loans in Last 3 months	f_m2_3_hl	0.98529
Maximum DPD for Non-ABC Housing Loans in Last 9 months	mxdpd_9_hl	0.98528
Number of Times Bucket 4+ for Non-ABC Housing Loans in Last 3 months	f_m4_3_hl	0.98528
Number of Times Bucket 6+ for Non-ABC Housing Loans in Last 9 months	f_m6_9_hl	0.98528
Number of Times Bucket 2+ for Non-ABC Housing Loans in Last 6 months	f_m2_6_hl	0.98528
Number of Times Bucket 5+ for Non-ABC Housing Loans in Last 3 months	f_m5_3_hl	0.98527
Number of Times Bucket 6+ for Non-ABC Housing Loans in Last 3 months	f_m6_3_hl	0.98527
Maximum DPD for Non-ABC Housing Loans in Last 12 months	mxdpd_12_hl	0.98527
Number of Times Bucket 3+ for Non-ABC Housing Loans in Last 9 months	f_m3_9_hl	0.98526
Number of Times Bucket 5+ for Non-ABC Housing Loans in Last 12 months	f_m5_12_hl	0.98525
Number of Times Bucket 6+ for Non-ABC Housing Loans in Last 12 months	f_m6_12_hl	0.98525
Number of Times Bucket 4+ for Non-ABC Housing Loans in Last 12 months	f_m4_12_hl	0.98525
Number of Times Bucket 1+ for Non-ABC Housing Loans in Last 3 months	f_m1_3_hl	0.98523
Number of Times Bucket 3+ for Non-ABC Housing Loans in Last 12 months	f_m3_12_hl	0.98520
Number of Times Bucket 2+ for Non-ABC Housing Loans in Last 9 months	f_m2_9_hl	0.98517
Number of Times Bucket 1+ for Non-ABC Housing Loans in Last 6 months	f_m1_6_hl	0.98513
Number of Times Bucket 2+ for Non-ABC Housing Loans in Last 12 months	f_m2_12_hl	0.98508
Number of Times Bucket 1+ for Non-ABC Housing Loans in Last 9 months	f_m1_9_hl	0.98489
Number of Times Bucket 1+ for Non-ABC Housing Loans in Last 12 months	f_m1_12_hl	0.98473

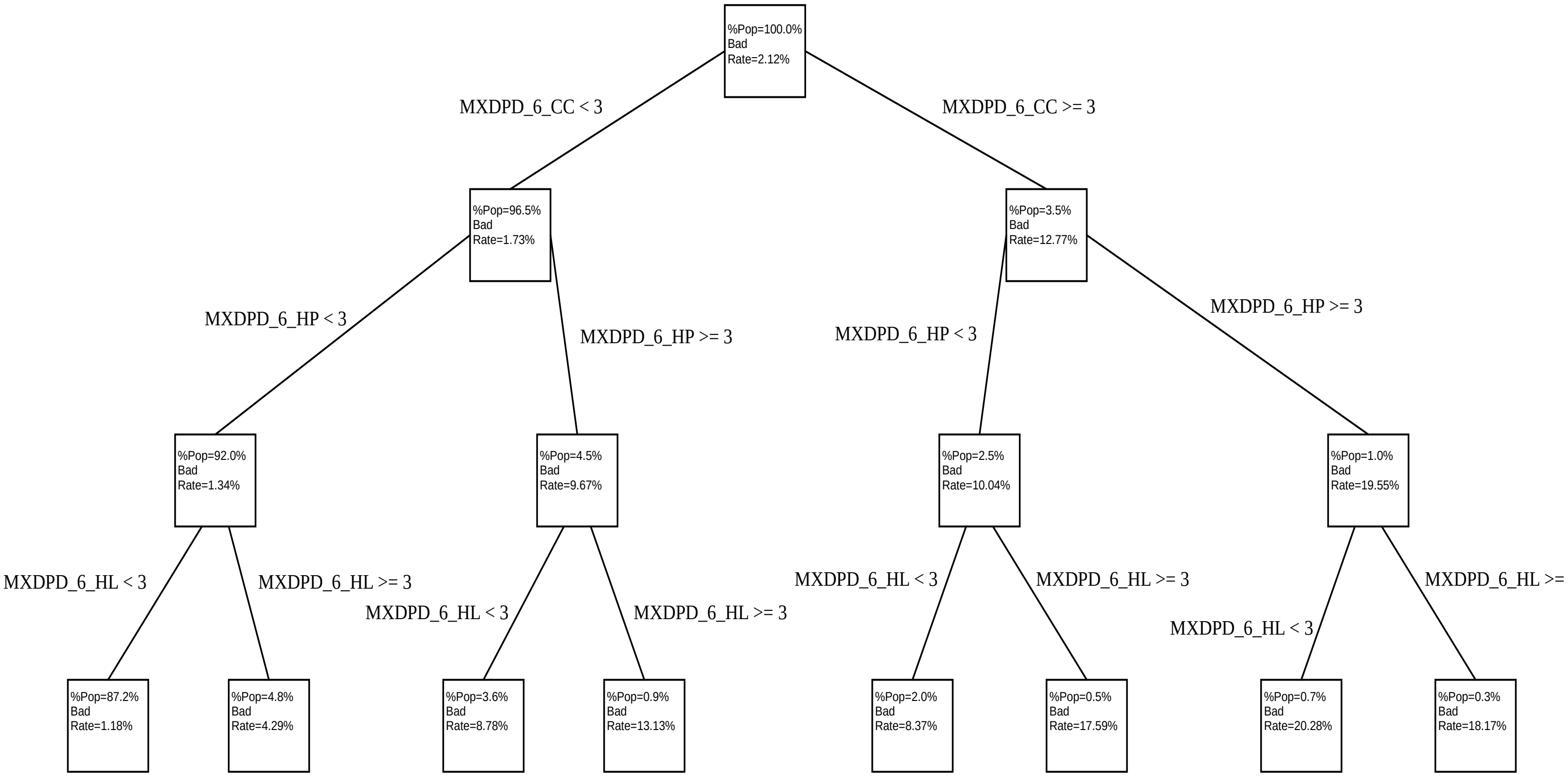
Factor Analysis – Important Factors – Case 2 – Total 1000 variables

Equivalent variables in Factor 4:

Number of Times Bucket 3+ for Non-ABC Other Facilities in Last 6 months	f_m3_6_ot	0.98351
Latest Maximum DPD for Non-ABC Other Facilities	f_ldpd_ot	0.98351
Number of Times Bucket 4+ for Non-ABC Other Facilities in Last 9 months	f_m4_9_ot	0.98350
Number of Times Bucket 4+ for Non-ABC Other Facilities in Last 6 months	f_m4_6_ot	0.98350
Number of Times Bucket 3+ for Non-ABC Other Facilities in Last 9 months	f_m3_9_ot	0.98350
Number of Times Bucket 5+ for Non-ABC Other Facilities in Last 9 months	f_m5_9_ot	0.98350
Maximum DPD for Non-ABC Other Facilities in Last 3 months	mxdpd_3_ot	0.98350
Number of Times Bucket 5+ for Non-ABC Other Facilities in Last 6 months	f_m5_6_ot	0.98350
Maximum DPD for Non-ABC Other Facilities in Last 9 months	mxdpd_9_ot	0.98350
Maximum DPD for Non-ABC Other Facilities in Last 6 months	mxdpd_6_ot	0.98349
Number of Times Bucket 2+ for Non-ABC Other Facilities in Last 6 months	f_m2_6_ot	0.98349
Maximum DPD for Non-ABC Other Facilities in Last 12 months	mxdpd_12_ot	0.98349
Number of Times Bucket 4+ for Non-ABC Other Facilities in Last 12 months	f_m4_12_ot	0.98349
Number of Times Bucket 6+ for Non-ABC Other Facilities in Last 9 months	f_m6_9_ot	0.98349
Number of Times Bucket 6+ for Non-ABC Other Facilities in Last 6 months	f_m6_6_ot	0.98349
Number of Times Bucket 5+ for Non-ABC Other Facilities in Last 12 months	f_m5_12_ot	0.98349
Number of Times Bucket 3+ for Non-ABC Other Facilities in Last 12 months	f_m3_12_ot	0.98348
Number of Times Bucket 6+ for Non-ABC Other Facilities in Last 12 months	f_m6_12_ot	0.98348
Number of Times Bucket 2+ for Non-ABC Other Facilities in Last 3 months	f_m2_3_ot	0.98347
Number of Times Bucket 3+ for Non-ABC Other Facilities in Last 3 months	f_m3_3_ot	0.98347
Number of Times Bucket 4+ for Non-ABC Other Facilities in Last 3 months	f_m4_3_ot	0.98346
Number of Times Bucket 5+ for Non-ABC Other Facilities in Last 3 months	f_m5_3_ot	0.98346
Number of Times Bucket 1+ for Non-ABC Other Facilities in Last 3 months	f_m1_3_ot	0.98346
Number of Times Bucket 2+ for Non-ABC Other Facilities in Last 9 months	f_m2_9_ot	0.98346
Number of Times Bucket 6+ for Non-ABC Other Facilities in Last 3 months	f_m6_3_ot	0.98346
Number of Times Bucket 1+ for Non-ABC Other Facilities in Last 6 months	f_m1_6_ot	0.98344
Number of Times Bucket 2+ for Non-ABC Other Facilities in Last 12 months	f_m2_12_ot	0.98342
Number of Times Bucket 1+ for Non-ABC Other Facilities in Last 9 months	f_m1_9_ot	0.98335
Number of Times Bucket 1+ for Non-ABC Other Facilities in Last 12 months	f_m1_12_ot	0.98327

Final Selection of Variables for Rule Generation – Case 2

Policy Rule



SUMMARY :

ASSUMPTIONS & TIPS FOR APPLYING FACTOR ANALYSIS IN PRACTICE

Factor Analysis Steps

1. Test assumptions

2. Select type of analysis

3. Determine no. of factors
   - Eigen Values, Scree Plot, % variance explained

4. Select items
   - check factor loadings to identify which items belong in which factor; drop items one by one; repeat 2-4

5. Label and define factors

6. Examine correlations amongst factors

7. Analyse internal reliability

8. Compute composite scores

Assumptions

- Sample size
- Levels of measurement
- Normality (sometimes difficult to check)
- Linearity (sometimes difficult to check)
- Outliers
- Factorability

Sample Size

Some heuristic guidelines:

Min.: *N* > 50 cases per variable

- e.g., 20 variables, should have > 1000 cases (1:50)

Ideal: *N* > 100 cases per variable

- e.g., 20 variables, ideally have > 2000 cases (1:100)

Total *N* > 2000 preferable

Level of Measurement

All variables must be suitable for correlational analysis

- i.e., they should be interval/continuous/numeric data or at least Likert Scale data with several interval levels.

Normality

Factor Analysis is generally robust to minor violation of assumptions of normality.

If the variables are normally distributed then the solution is **enhanced**.

Linearity

Factor Analysis is based on correlations between variables.

Therefore, it is important to check if there are linear relations amongst the variables (i.e., check scatterplots)

Outliers

Factor Analysis is sensitive to outlying cases.

- Bivariate outliers
  (e.g., check scatterplots)
- Multivariate outliers
  (e.g., Mahalanobis' distance)

Identify outliers, then remove or transform

Interpretation of Communalities

High communalities (> .5): Extracted factors explain most of the variance in the variables being analysed

Low communalities (< .5): A variable has considerable variance unexplained by the extracted factors

- May then need to extract MORE factors to explain the variance or remove these items

Explained Variance

A good factor solution is one that explains the most variance with the fewest factors.

Realistically, researchers are happy with 50-75% of the variance explained.

Eigen Values (EVs)

Each factor has an EV which indicates the amount of variance each factor accounts for.

EVs for successive factors have lower values.

Rule of thumb: Eigen values over 1 are ‘stable’ (Kaiser's criterion).

EVs can also be expressed as % s.

Total of all EVs is the number of variables. Each variable contributes a variance of one. EVs are then allocated to factors according to amount of variance explained.

Scree Plot

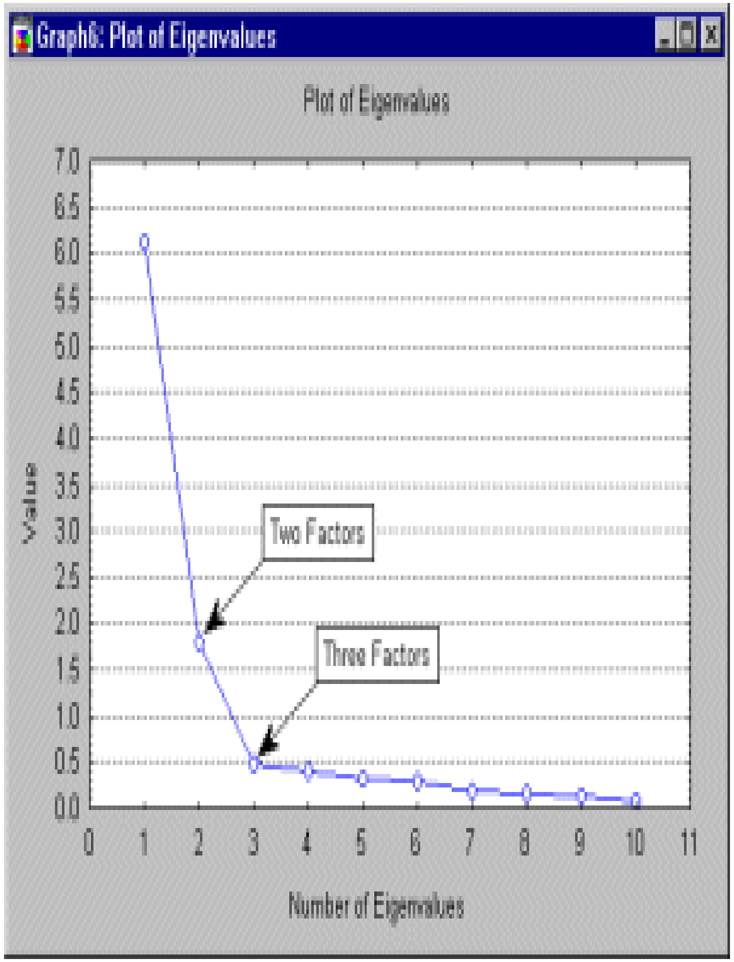
A line graph of EVs.

Depicts amount of variance explained by each factor.

Cut-off: Look for where additional factors fail to add appreciably to the cumulative explained variance.

1st factor explains the most variance.

Last factor explains the least amount of variance.



How Many Factors?

A subjective process ... Seek to explain maximum variance using fewest factors, considering:

- Theory – what is predicted/expected?
- Eigen Values > 1? (Kaiser's criterion)
- Scree Plot – where does it drop off?
- Interpretability of last factor?
- Try several different solutions?
  (consider FA type, rotation, # of factors)
- Factors must be able to be meaningfully interpreted & make theoretical sense?

How Many Factors

Aim for 50-75% of variance explained by 1/4 to 1/3 as many factors as variables/items.

Stop extracting factors when they no longer represent useful/meaningful clusters of variables.

Keep checking/clarifying the meaning of each factor – make sure to examine the wording of each item.

Initial solution: Un-rotated factor structure

Factor loadings (FLs) indicate relative importance of each item to each factor.

- In the initial solution, each factor tries “selfishly” to grab maximum unexplained variance.
- All variables will tend to load strongly on the 1st factor

Factors are weighted combinations of original raw variables

Initial solution: Un-rotated factor structure

1st factor extracted:

- Best possible line of best fit through the original variables
- Seeks to explain lion's share of all variance
- A single factor, best summary of the variance in the whole set of items

Each subsequent factor tries to explain the maximum possible amount of remaining unexplained variance.

- Second factor is orthogonal to first factor - seeks to maximise its own eigen value (i.e., tries to gobble up as much of the remaining unexplained variance as possible)

Initial solution: Un-rotated factor structure

Seldom see a simple unrotated factor structure

Many variables load on 2 or more factors

Some variables may not load highly on any factors (check: low communality)

Until the Factor Loadings are rotated, they are difficult to interpret.

Rotation of the Factor Loadings matrix helps to find a more interpretable factor structure.

Why Rotate a Factor Loading Matrix?

After rotation, the vectors (lines of best fit) are rearranged to optimally go through clusters of shared variance

Then the Factor Loadings and the factor they represent can be more readily interpreted.

A rotated factor structure is simpler & more easily interpretable

- each variable loads strongly on only one factor
- each factor shows at least 3 strong loadings
- all loading are either strong or weak, no intermediate loadings

Interpretability

It is dangerous to be driven by factor loadings only – think carefully - be guided by theory and common sense in selecting factor structure.

You must be able to understand and interpret a factor if you're going to extract it.

However, watch out for 'seeing what you want to see' when evidence might suggest a different, better solution.

There may be more than one good solution! e.g., in personality

- 2 factor model
- 5 factor model
- 16 factor model

How many Items per Factor?

Bare min. = 2

Recommended min. = 3

Max. = unlimited

More items:

- increase reliability
- Increase 'roundedness‘
- Law of diminishing returns

Typically = 4 to 10 is reasonable

How do I eliminate items?

A subjective process; Consider:

- Size of main loading (min = .4)
- Size of cross loadings (max = .3?)
- Meaning of item (face validity)
- Contribution it makes to the factor
- Eliminate 1 variable at a time, then re-run, before deciding which/if any items to eliminate next
- Number of items already in the factor

Factor Loadings & Item Selection

Comrey & Lee (1992):

- loadings > .70 - excellent
- > .63 - very good
- > .55 - good
- > .45 - fair
- > .32 – poor
- < .32 – very poor

Cut-off for acceptable loadings:

- Look for gap in loadings
  (e.g., .8, .7, .6, .3, .2)

Choose cut-off because factors can be interpreted above but not below cut-off

Factor Analysis in Practice

To find a good solution, consider each combination of:

- PC-Varimax
- PC-Oblimin

Apply the above methods to a range of possible/likely factors, e.g., for 2, 3, 4, 5, 6, and 7 factors

Eliminate poor items one at a time, retesting the possible solutions

Check factor structure across sub-groups (e.g., gender) if there is sufficient data

You will probably come up with a different solution from someone else!

A PRACTICAL EXAMPLE OF FACTOR ANALYSIS USING SPSS

Practical Example of Factor Analysis

- We have a dataset describing various attributes (X_i) of a set of Cars

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	manufact	String	13	0	Manufacturer	None	None	13	Left	Nominal	Input
2	model	String	17	0	Model	None	None	17	Left	Nominal	Input
3	sales	Numeric	11	3	Sales in thousa...	None	None	8	Right	Scale	Input
4	resale	Numeric	11	3	4-year resale va...	None	None	8	Right	Scale	Input
5	type	Numeric	11	0	Vehicle type	{0, Automob...	None	8	Right	Ordinal	Input
6	price	Numeric	11	3	Price in thousa...	None	None	8	Right	Scale	Input
7	engine_s	Numeric	11	1	Engine size	None	None	8	Right	Scale	Input
8	horsepow	Numeric	11	0	Horsepower	None	None	8	Right	Scale	Input
9	wheelbas	Numeric	11	1	Wheelbase	None	None	8	Right	Scale	Input
10	width	Numeric	11	1	Width	None	None	8	Right	Scale	Input
11	length	Numeric	11	1	Length	None	None	8	Right	Scale	Input
12	curb_wgt	Numeric	11	3	Curb weight	None	None	8	Right	Scale	Input
13	fuel_cap	Numeric	11	1	Fuel capacity	None	None	8	Right	Scale	Input
14	mpg	Numeric	11	0	Fuel efficiency	None	None	8	Right	Scale	Input

- Obviously these attributes are correlated with each other, so is it possible to construct a smaller set of uncorrelated components?

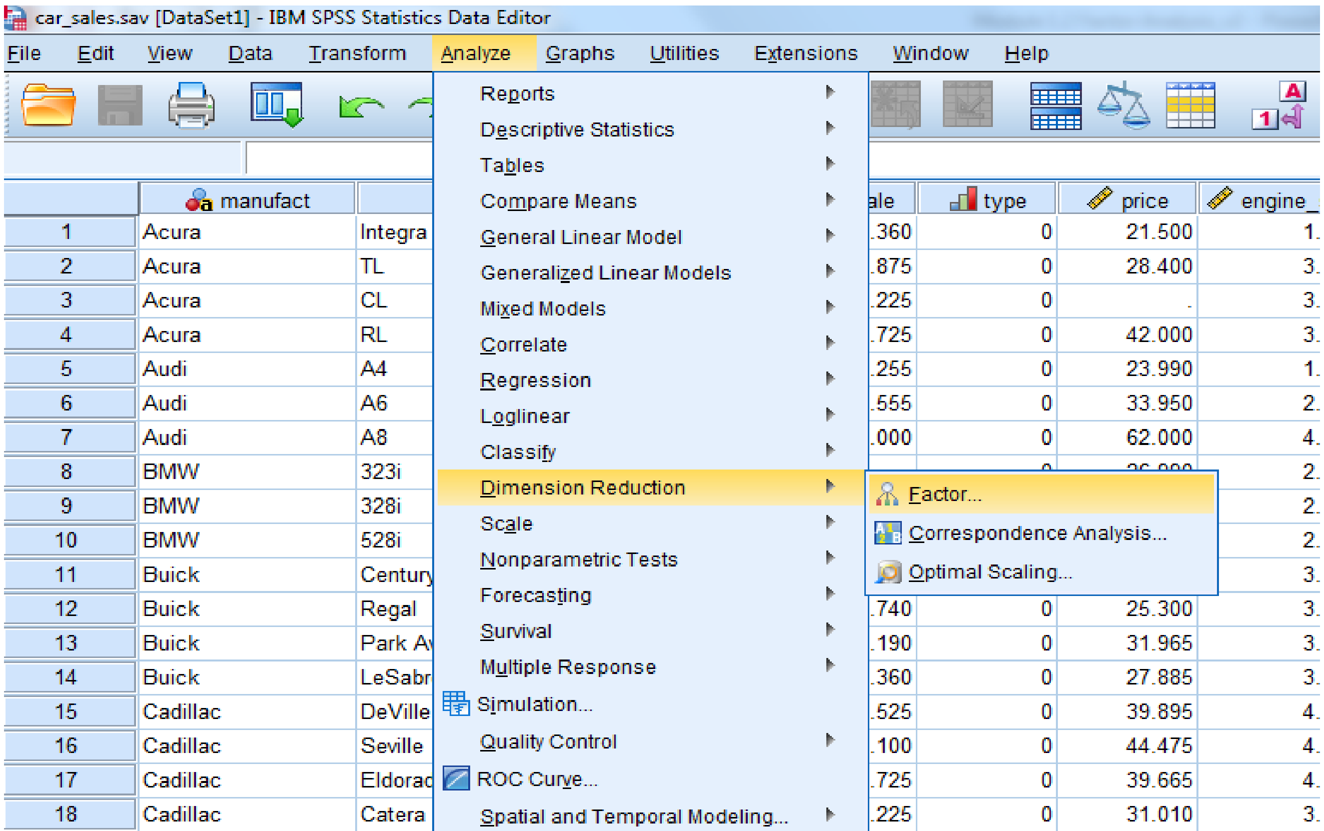
Practical Example of Factor Analysis

Sample Data

	manufact	model	sales	resale	type	price	engine_s	horsepow	wheelbas	width	length	curb_wgt	fuel_cap
1	Acura	Integra	16.919	16.360	0	21.500	1.8	140	101.2	67.3	172.4	2.639	13.2
2	Acura	TL	39.384	19.875	0	28.400	3.2	225	108.1	70.3	192.9	3.517	17.2
3	Acura	CL	14.114	18.225	0	.	3.2	225	106.9	70.6	192.0	3.470	17.2
4	Acura	RL	8.588	29.725	0	42.000	3.5	210	114.6	71.4	196.6	3.850	18.0
5	Audi	A4	20.397	22.255	0	23.990	1.8	150	102.6	68.2	178.0	2.998	16.4
6	Audi	A6	18.780	23.555	0	33.950	2.8	200	108.7	76.1	192.0	3.561	18.5
7	Audi	A8	1.380	39.000	0	62.000	4.2	310	113.0	74.0	198.2	3.902	23.7
8	BMW	323i	19.747	.	0	26.990	2.5	170	107.3	68.4	176.0	3.179	16.6
9	BMW	328i	9.231	28.675	0	33.400	2.8	193	107.3	68.5	176.0	3.197	16.6
10	BMW	528i	17.527	36.125	0	38.900	2.8	193	111.4	70.9	188.0	3.472	18.5
11	Buick	Century	91.561	12.475	0	21.975	3.1	175	109.0	72.7	194.6	3.368	17.5
12	Buick	Regal	39.350	13.740	0	25.300	3.8	240	109.0	72.7	196.2	3.543	17.5
13	Buick	Park Avenue	27.851	20.190	0	31.965	3.8	205	113.8	74.7	206.8	3.778	18.5
14	Buick	LeSabre	83.257	13.360	0	27.885	3.8	205	112.2	73.5	200.0	3.591	17.5
15	Cadillac	DeVille	63.729	22.525	0	39.895	4.6	275	115.3	74.5	207.2	3.978	18.5
16	Cadillac	Seville	15.943	27.100	0	44.475	4.6	275	112.2	75.0	201.0	.	18.5
17	Cadillac	Eldorado	6.536	25.725	0	39.665	4.6	275	108.0	75.5	200.6	3.843	19.0
18	Cadillac	Catera	11.185	18.225	0	31.010	3.0	200	107.4	70.3	194.8	3.770	18.0

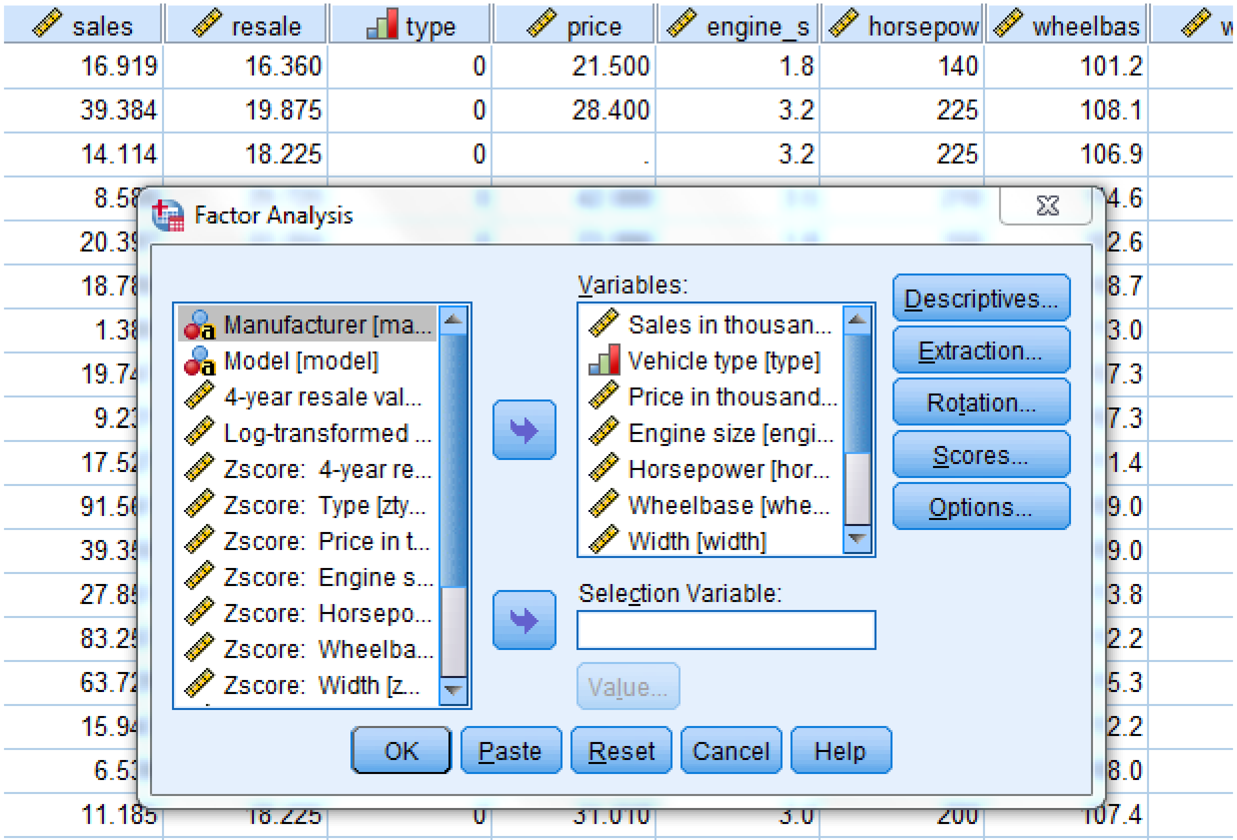
Practical Example of Factor Analysis

- Access Car_sales SPSS Dataset, then access Analyze -> Dimension Reduction - > Factor



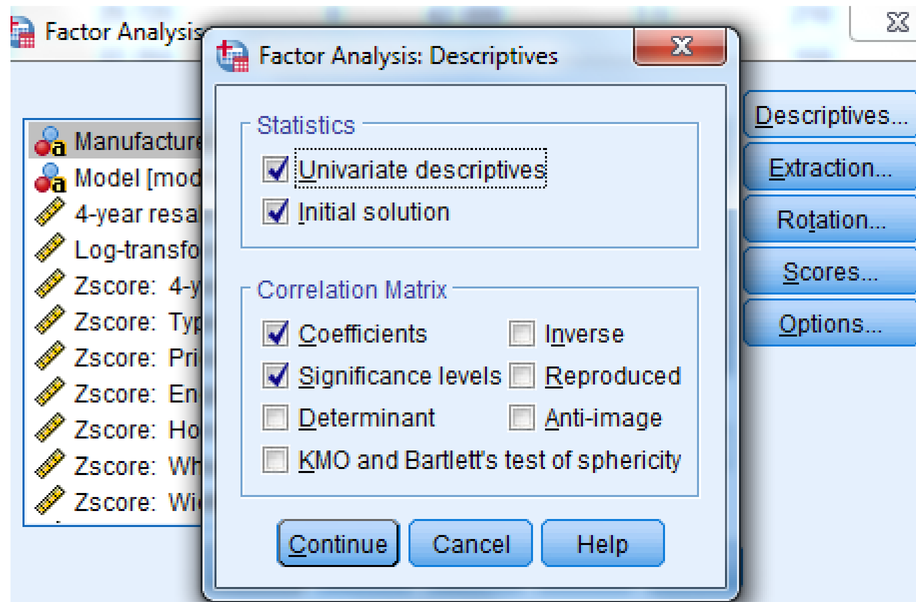
Practical Example of Factor Analysis

- Then select variables → Only continuous or scale variable

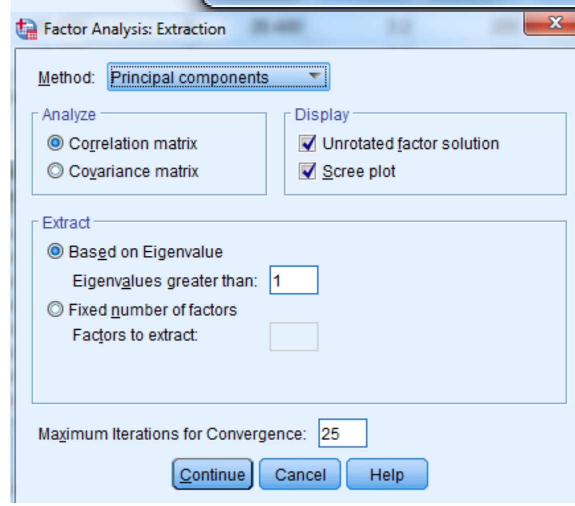


Practical Example of Factor Analysis

- Go to Descriptives and select necessary options



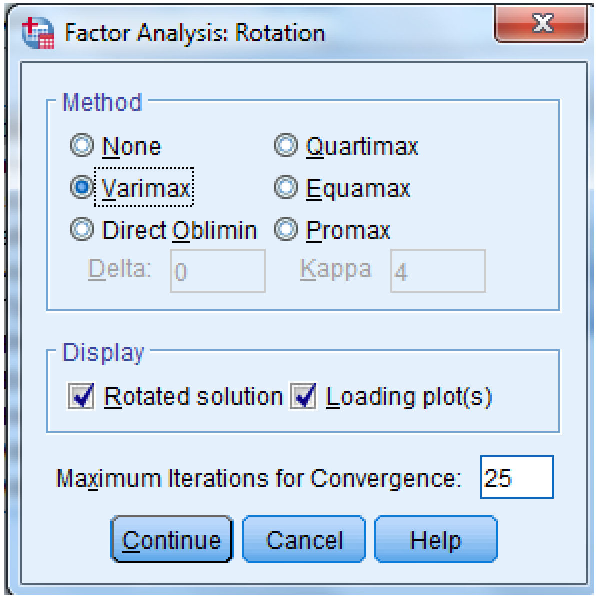
- Select Extraction method

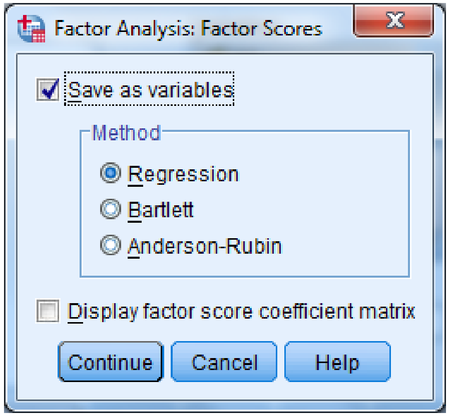


Practical Example of Factor Analysis

- Go to the *Rotation* Window and make necessary choice

- Go to the *Scores* Window
  - Tick the *Save as variables* Box
  - *Method* should be *Regression*
- Then press *OK* on the factor analysis box





Factor Analysis - Output

Factor Analysis

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
Vehicle type	.26	.442	152
Price in thousands	27.33182	14.418669	152
Engine size	3.049	1.0498	152
Horsepower	184.81	56.823	152
Wheelbase	107.414	7.7178	152
Width	71.089	3.4647	152
Length	187.059	13.4712	152
Curb weight	3.37618	.636593	152
Fuel capacity	17.959	3.9376	152
Fuel efficiency	23.84	4.305	152

Correlation Matrix

		Vehicle type	Price in thousands	Engine size	Horsepower	Wheelbase	W
Correlation	Vehicle type	1.000	-.042	.269	.017	.397	
	Price in thousands	-.042	1.000	.624	.841	.108	
	Engine size	.269	.624	1.000	.837	.473	
	Horsepower	.017	.841	.837	1.000	.282	
	Wheelbase	.397	.108	.473	.282	1.000	
	Width	.260	.328	.692	.535	.681	
	Length	.150	.155	.542	.385	.840	
	Curb weight	.526	.527	.761	.611	.651	
	Fuel capacity	.599	.424	.667	.505	.657	
	Fuel efficiency	-.577	-.492	-.737	-.616	-.497	
Sig. (1-tailed)	Vehicle type		.303	.000	.419	.000	
	Price in thousands	.303		.000	.000	.092	
	Engine size	.000	.000		.000	.000	
	Horsepower	.419	.000	.000		.000	
	Wheelbase	.000	.092	.000	.000		
	Width	.001	.000	.000	.000	.000	
	Length	.033	.028	.000	.000	.000	
	Curb weight	.000	.000	.000	.000	.000	
	Fuel capacity	.000	.000	.000	.000	.000	
	Fuel efficiency	.000	.000	.000	.000	.000	

Factor Analysis - Output

- The initial output shows the major principal components created and % of variance they explained

Total Variance Explained

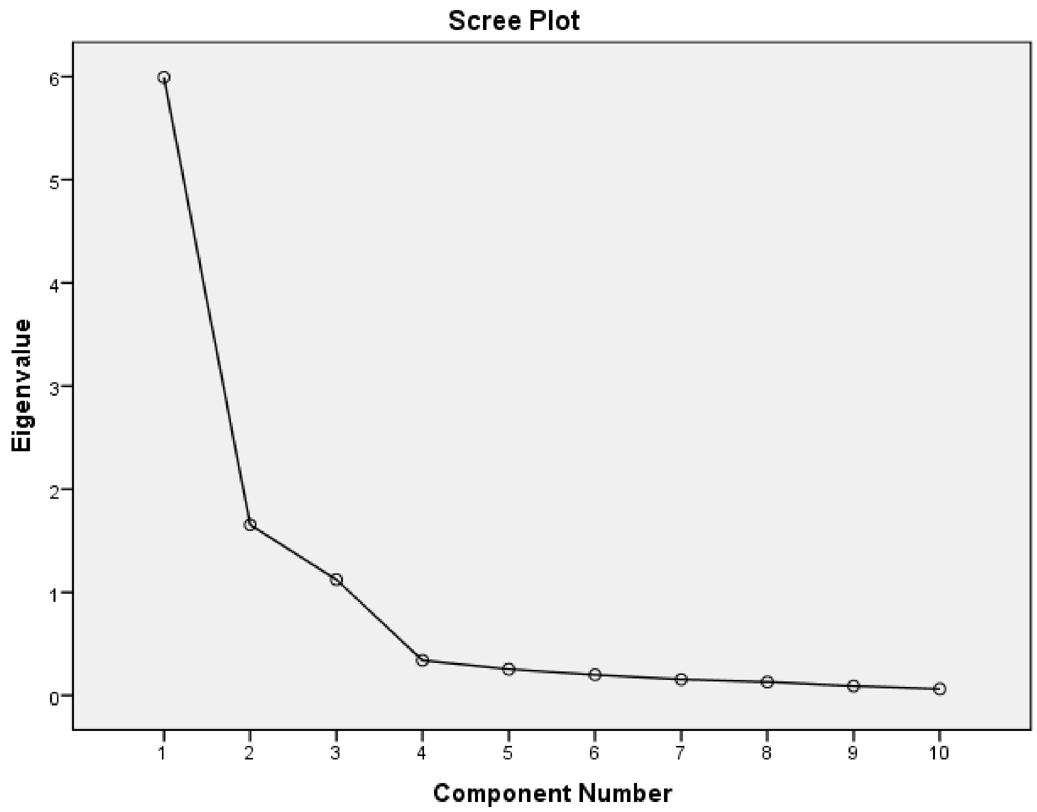
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.994	59.938	59.938	5.994	59.938	59.938	3.220	32.199	32.199
2	1.654	16.545	76.482	1.654	16.545	76.482	3.134	31.344	63.543
3	1.123	11.227	87.709	1.123	11.227	87.709	2.417	24.166	87.709
4	.339	3.389	91.098						
5	.254	2.541	93.640						
6	.199	1.994	95.633						
7	.155	1.547	97.181						
8	.130	1.299	98.480						
9	.091	.905	99.385						
10	.061	.615	100.000						

Extraction Method: Principal Component Analysis.

- As you can see the first three principal components explain 83% of the variance, and their eigenvalues are all >1

Factor Analysis - Output

- The Scree plot shows the drop in successive eigenvalues



Factor Analysis - Output

- And the weights for each variable in each principal component are shown here

Component Matrix[a]

	Component		
	1	2	3
Vehicle type	.471	.533	-.651
Price in thousands	.580	-.729	-.092
Engine size	.871	-.290	.018
Horsepower	.740	-.618	.058
Wheelbase	.732	.480	.340
Width	.821	.114	.298
Length	.719	.304	.556
Curb weight	.934	.063	-.121
Fuel capacity	.885	.184	-.210
Fuel efficiency	-.863	.004	.339

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

Rotated Component Matrix[a]

	Component		
	1	2	3
Vehicle type	-.101	.095	.954
Price in thousands	.935	-.003	.041
Engine size	.753	.436	.292
Horsepower	.933	.242	.056
Wheelbase	.036	.884	.314
Width	.384	.759	.231
Length	.155	.943	.069
Curb weight	.519	.533	.581
Fuel capacity	.398	.495	.676
Fuel efficiency	-.543	-.318	-.681

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 4 iterations.

The Business Problem

Financial institutions across the world are increasingly offering customers mobile banking and finance services.

A particular financial institution would like to understand what drives smart phone users to use financial applications.

Further, stated importance versus derived importance identifies the key drivers of financial application usage. Smartphone users who use financial applications are profiled to enhance target marketing.

Hidden opportunities are identified and this opens the doors for organizations to be innovative and to make strategic decisions by differentiating themselves and thereby gaining a competitive advantage in the market. This financial institution would also like to explore new ‘Concepts’ and new product ideas.

Guidelines for Factor Analysis

The first thing to do when conducting a Factor Analysis is to do look at the summary statistics of all the relevant variables and the inter-Correlations between variables.

If the survey questions measure the same underlying dimensions the we would expect them to correlate with each other (because they are measuring the same thing).

If we find any variables that does not correlate with any other variables (or very few variables) then we should exclude these variables before the Factor Analysis is run.

In general over 300 cases is probably adequate but the communalities after extraction should probably be above 0.5.

Guidelines for Factor Analysis

Although mild Multi-collinearity is not a problem for Factor Analysis. It is important to avoid extreme Multi-collinearity (i.e. variables that are very highly correlated) and Singularity (variables that are perfectly correlated).

Singularity causes problems in Factor Analysis as it becomes impossible to determine the unique contribution to a Factor of the variables that are highly correlated.

At this stage we look to eliminating any variable that does not correlate with any other variable or we exclude any variable that is highly correlated with another variable.

As well as looking at interrelations, all variables should be roughly normally distributed and are measured at an interval level (which rScales are wrongly assumed to be!)

The assumption of normality is important only if you wish to generalize the results of your analysis beyond the sample collected.

Readings on Factor Analysis

- Book Reference : **Subhash Sharma, Applied Multivariate Techniques, Wiley**
- Alternative Perceptual Mapping Techniques:Relative Accuracy and Usefulness, *Journal of Marketing Research,* JOHN R. HAUSER and FRANK S. KOPPELMAN, Vol. XVI (November 1979), 495-506
- An Easy Guide to Factor Analysis by Paul Kline
- Making Sense of Factor Analysis: The Use of Factor Analysis for Instrument Development in Health Care Research by Marjorie A. Pett, Nancy R. Lackey and John J. Sullivan
- Factor Analysis at 100: Historical Developments and Future Directions edited by Robert Cudeck and Robert C. MacCallum
- Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications by Bruce Thompson
- A First Course in Factor Analysis by Andrew L. Comrey
- Modern Factor Analysis, Second Edition, Revised by Harry H. Harman
- Confirmatory Factor Analysis for Applied Research by Timothy A. Brown.

Thank You!