Introduction to Statistical Concepts & Tools For Factor Analysis

By

Dr. Rajeev Pandey

Professor & Head

Department of Statistics

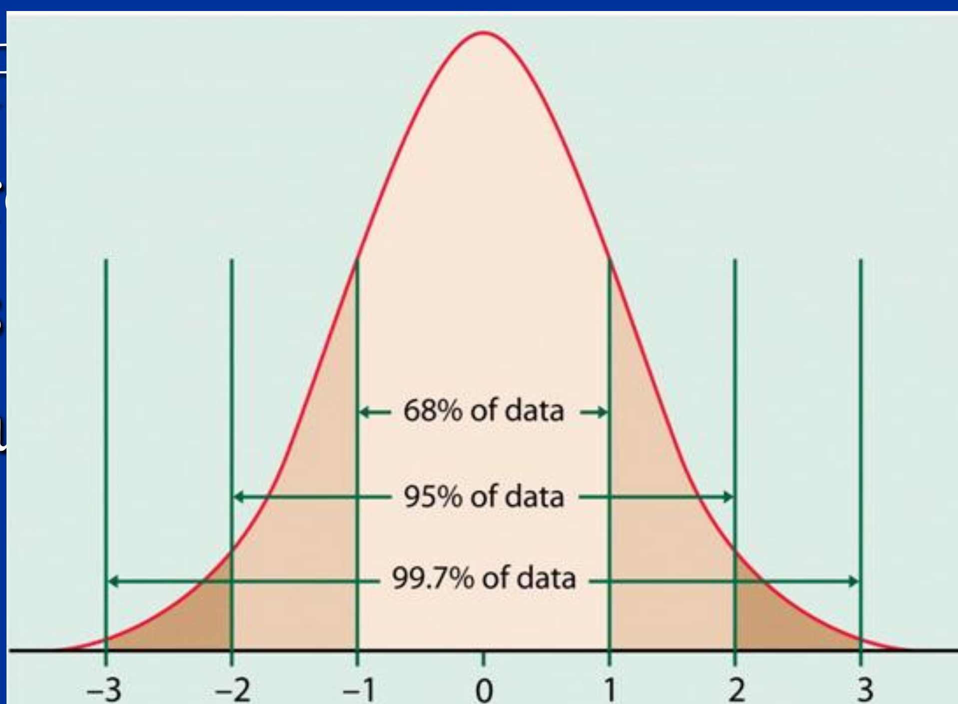
University of Lucknow

pandey_rajeev@lkouniv.ac.in prof.rajeevlu@gmail.com

Web: http://dacg.in

Role of Normality

- Many statistical methods require that the numeric variables we are working with have an approximate **normal distribution**.
- F ts, and r ire in some s ables are a stributed.



Standardized normal distribution with empirical rule percentages.

Tools for Assessing Normality

- Histogram and Boxplot
- Normal Quantile Plot
  (also called Normal Probability Plot)
- Goodness of Fit Tests

  Shapiro-Wilk Test (JMP)

  Kolmogorov-Smirnov Test (SPSS)

  Anderson-Darling Test (MINITAB)

Factor Analysis

1) **Overview**
2) **Basic Concept**
3) **Factor Analysis Model**
4) **Statistics Associated with Factor Analysis**

5) Conducting Factor Analysis

1. **Problem Formulation**
2. **Construction of the Correlation Matrix**
3. **Method of Factor Analysis**
4. **Number of of Factors**
5. **Rotation of Factors**
6. **Interpretation of Factors**
7. **Factor Scores**
8. **Model Fit**

Factor Analysis

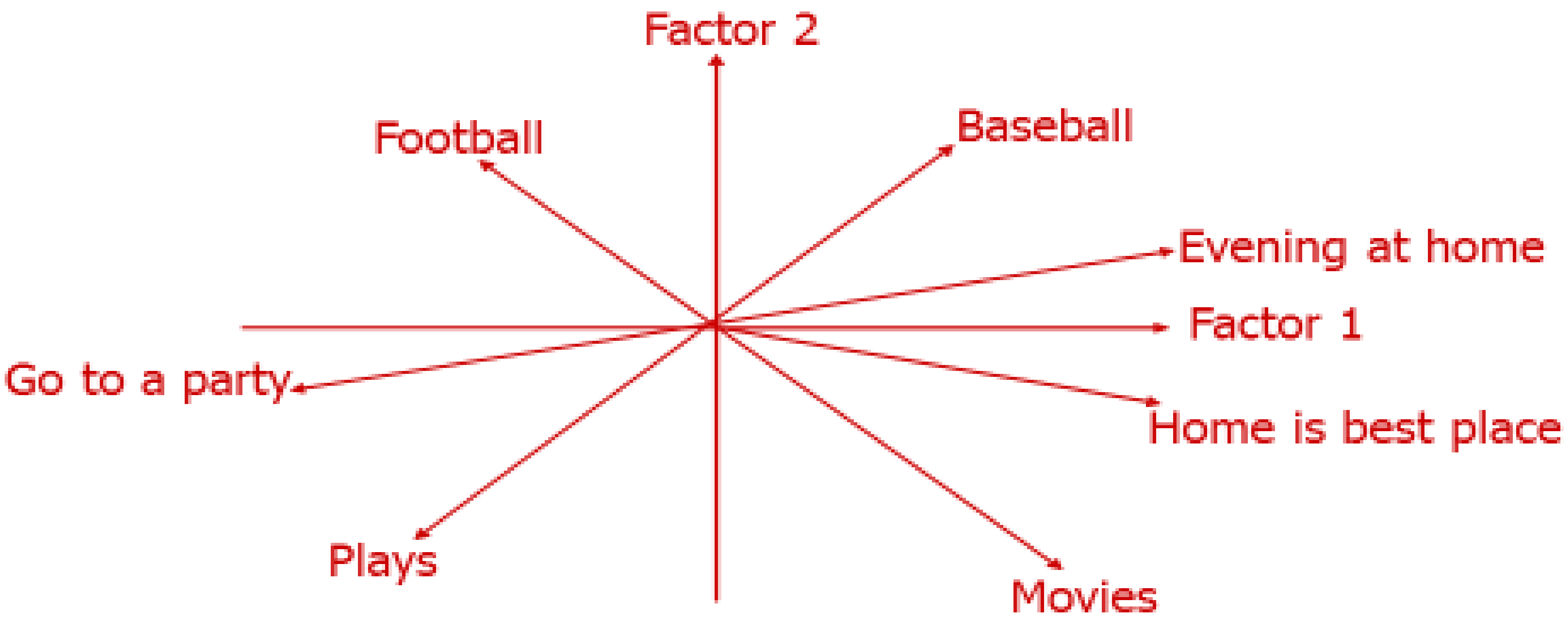
- **Factor analysis** is a class of procedures used for data reduction and summarization.
- It is an **interdependence technique:** no distinction between dependent and independent variables.

- Factor analysis is used:
  - To identify underlying dimensions, or **factors**, that explain the correlations among a set of variables.
  - To identify a new, smaller, set of uncorrelated variables to replace the original set of correlated variables.

Factor Analysis - Example

Factors Underlying Selected Psychographics and Lifestyles

Fig. 19.1



Factor Analysis Model

Each variable is expressed as a linear combination of factors. The factors are some common factors plus a unique factor. The factor model is represented as:

$$X_i = A_{i1}F_1 + A_{i2}F_2 + A_{i3}F_3 + \ldots + A_{im}F_m + V_iU_i$$

where

X_i = i th standardized variable

A_{ij} = standardized mult reg coeff of var i on common factor j

F_j = common factor j

V_i = standardized reg coeff of var i on unique factor i

U_i = the unique factor for variable i

m = number of common factors

Factor Analysis Model

- The first set of weights (factor score coefficients) are chosen so that the first factor explains the largest portion of the total variance.
- Then a second set of weights can be selected, so that the second factor explains most of the residual variance, subject to being uncorrelated with the first factor.
- This same principle applies for selecting additional weights for the additional factors.

Factor Analysis Model

The common factors themselves can be expressed as linear combinations of the observed variables.

$$F_i = W_{i1}X_1 + W_{i2}X_2 + W_{i3}X_3 + \ldots + W_{ik}X_k$$

Where:

F_i = estimate of *i* th factor

W_i= weight or factor score coefficient

k = number of variables

Statistics Associated with Factor Analysis

- **Bartlett's test of sphericity**. Bartlett's test of sphericity is used to test the hypothesis that the variables are uncorrelated in the population (i.e., the population corr matrix is an identity matrix)

- **Correlation matrix**. A correlation matrix is a lower triangle matrix showing the simple correlations, *r*, between all possible pairs of variables included in the analysis. The diagonal elements are all 1.

Statistics Associated with Factor Analysis

- **Communality**. Amount of variance a variable shares with all the other variables. This is the proportion of variance explained by the common factors.
- **Eigenvalue**. Represents the total variance explained by each factor.
- **Factor loadings**. Correlations between the variables and the factors.
- **Factor matrix**. A factor matrix contains the factor loadings of all the variables on all the factors

Statistics Associated with Factor Analysis

- **Factor scores**. Factor scores are composite scores estimated for each respondent on the derived factors.
- **Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy**. Used to examine the appropriateness of factor analysis. High values (between 0.5 and 1.0) indicate appropriateness. Values below 0.5 imply not.
- **Percentage of variance**. The percentage of the total variance attributed to each factor.
- **Scree plot**. A scree plot is a plot of the Eigenvalues against the number of factors in order of extraction.

Example: Factor Analysis

- HATCO is a large industrial supplier
- A marketing research firm surveyed 100 HATCO customers, to investigate the customers' perceptions of HATCO
- The marketing research firm obtained data on 7 different variables from HATCO's customers
- Before doing further analysis, the mkt res firm ran a Factor Analysis to see if the data could be reduced

Example: Factor Analysis

- In a B2B situation, HATCO wanted to know the perceptions that its customers had about it
- The mktg res firm gathered data on 7 variables
  1. Delivery speed
  2. Price level
  3. Price flexibility
  4. Manufacturer's image
  5. Overall service
  6. Salesforce image
  7. Product quality
- Each var was measured on a 10 cm graphic rating scale

Conducting Factor Analysis

- Problem formulation
- Construction of the Correlation Matrix
- Method of Factor Analysis
- Determination of Number of Factors
- Rotation of Factors
- Interpretation of Factors
- Calculation of Factor Scores
- Determination of Model Fit

Formulate the Problem

- The objectives of factor analysis should be identified.

- The variables to be included in the factor analysis should be specified. The variables should be measured on an interval or ratio scale.

- An appropriate sample size should be used. As a rough guideline, there should be at least four or five times as many observations (sample size) as there are variables.

Construct the Correlation Matrix

- The analytical process is based on a matrix of correlations between the variables.
- If the Bartlett's test of sphericity is not rejected, then factor analysis is not appropriate.
- If the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy is small, then the correlations between pairs of variables cannot be explained by other variables and factor analysis may not be appropriate.

Determine the Method of Factor Analysis

- In **Principal components analysis**, the total variance in the data is considered.

  -Used to determine the min number of factors that will account for max variance in the data.

- In **Common factor analysis**, the factors are estimated based only on the common variance.

  -Communalities are inserted in the diagonal of the correlation matrix.

  -Used to identify the underlying dimensions and when the common variance is of interest.

Determine the Number of Factors

- **A Priori Determination.** Use prior knowledge.
- **Determination Based on Eigenvalues.** Only factors with Eigenvalues greater than 1.0 are retained.
- **Determination Based on Scree Plot.** A scree plot is a plot of the Eigenvalues against the number of factors in order of extraction. The point at which the scree begins denotes the true number of factors.
- **Determination Based on Percentage of Variance.**

Rotation of Factors

- Through rotation the factor matrix is transformed into a simpler one that is easier to interpret.

- After rotation each factor should have nonzero, or significant, loadings for only some of the variables. Each variable should have nonzero or significant loadings with only a few factors, if possible with only one.

- The rotation is called **orthogonal rotation** if the axes are maintained at right angles.

Rotation of Factors

- **Varimax procedure.** Axes maintained at right angles

  -Most common method for rotation.

  -An orthogonal method of rotation that minimizes the number of variables with high loadings on a factor.

  -Orthogonal rotation results in uncorrelated factors.

- **Oblique rotation.** Axes not maintained at right angles

  -Factors are correlated.

  -Oblique rotation should be used when factors in the population are likely to be strongly correlated.

Interpret Factors

- A factor can be interpreted in terms of the variables that load high on it.
- Another useful aid in interpretation is to plot the variables, using the factor loadings as coordinates. Variables at the end of an axis are those that have high loadings on only that factor, and hence describe the factor.

Calculate Factor Scores

The factor scores for the i th factor may be estimated as follows:

$$F_i = W_{i1} X_1 + W_{i2} X_2 + W_{i3} X_3 + \dots + W_{ik} X_k$$

Determine the Model Fit

- The correlations between the variables can be deduced from the estimated correlations between the variables and the factors.

- The differences between the observed correlations (in the input correlation matrix) and the reproduced correlations (estimated from the factor matrix) can be examined to determine model fit. These differences are called *residuals*.

Another Example of Factor Analysis

- To determine benefits from toothpaste
- Responses were obtained on 6 variables:

  V1: It is imp to buy toothpaste to prevent cavities

  V2: I like a toothpaste that gives shiny teeth

  V3: A toothpaste should strengthen your gums

  V4: I prefer a toothpaste that freshens breath

  V5: Prevention of tooth decay is not imp

  V6: The most imp consideration is attractive teeth

- Responses on a 7-pt scale (1=strongly disagree; 7=strongly agree)

Another Example of Factor Analysis

RESPONDENT NUMBER	V1	V2	V3	V4	V5	V6
1	7.00	3.00	6.00	4.00	2.00	4.00
2	1.00	3.00	2.00	4.00	5.00	4.00
3	6.00	2.00	7.00	4.00	1.00	3.00
4	4.00	5.00	4.00	6.00	2.00	5.00
5	1.00	2.00	2.00	3.00	6.00	2.00
6	6.00	3.00	6.00	4.00	2.00	4.00
7	5.00	3.00	6.00	3.00	4.00	3.00
8	6.00	4.00	7.00	4.00	1.00	4.00
9	3.00	4.00	2.00	3.00	6.00	3.00
10	2.00	6.00	2.00	6.00	7.00	6.00
11	6.00	4.00	7.00	3.00	2.00	3.00
12	2.00	3.00	1.00	4.00	5.00	4.00
13	7.00	2.00	6.00	4.00	1.00	3.00
14	4.00	6.00	4.00	5.00	3.00	6.00
15	1.00	3.00	2.00	2.00	6.00	4.00
16	6.00	4.00	6.00	3.00	3.00	4.00
17	5.00	3.00	6.00	3.00	3.00	4.00
18	7.00	3.00	7.00	4.00	1.00	4.00
19	2.00	4.00	3.00	3.00	6.00	3.00
20	3.00	5.00	3.00	6.00	4.00	6.00
21	1.00	3.00	2.00	3.00	5.00	3.00
22	5.00	4.00	5.00	4.00	2.00	4.00
23	2.00	2.00	1.00	5.00	4.00	4.00
24	4.00	6.00	4.00	6.00	4.00	7.00
25	6.00	5.00	4.00	2.00	1.00	4.00
26	3.00	5.00	4.00	6.00	4.00	7.00
27	4.00	4.00	7.00	2.00	2.00	5.00
28	3.00	7.00	2.00	6.00	4.00	3.00
29	4.00	6.00	3.00	7.00	2.00	7.00
30	2.00	3.00	2.00	4.00	7.00	2.00

Correlation Matrix

Variables	V1	V2	V3	V4	V5	V6
V1	1.000					
V2	-0.530	1.000				
V3	0.873	-0.155	1.000			
V4	-0.086	0.572	-0.248	1.000		
V5	-0.858	0.020	-0.778	-0.007	1.000	
V6	0.004	0.640	-0.018	0.640	-0.136	1.000

Results of Principal Components Analysis

Bartlett's Test
Apprx. chi-square=111.3, df=15, significance=0.00
Kaiser-Meyer-Olkin msa=0.660

Communalities

Variables	Initial	Extraction
V1	1.000	0.926
V2	1.000	0.723
V3	1.000	0.894
V4	1.000	0.739
V5	1.000	0.878
V6	1.000	0.790

Initial Eigen values

Factor	Eigen value	% of variance	Cumulat. %
1	2.731	45.520	45.520
2	2.218	36.969	82.488
3	0.442	7.360	89.848
4	0.341	5.688	95.536
5	0.183	3.044	98.580
6	0.085	1.420	100.000

Results of Principal Components Analysis

Extraction Sums of Squared Loadings

Factor	Eigen value	% of variance	Cumulat. %
1	2.731	45.520	45.520
2	2.218	36.969	82.488

Factor Matrix

Variables	Factor 1	Factor 2
V1	0.928	0.253
V2	-0.301	0.795
V3	0.936	0.131
V4	-0.342	0.789
V5	-0.869	-0.351
V6	-0.177	0.871

Rotation Sums of Squared Loadings

Factor	Eigenvalue	% of variance	Cumulat. %
1	2.688	44.802	44.802
2	2.261	37.687	82.488

Results of Principal Components Analysis

Rotated Factor Matrix

Variables	Factor 1	Factor 2
V1	0.962	-0.027
V2	-0.057	0.848
V3	0.934	-0.146
V4	-0.098	0.845
V5	-0.933	-0.084
V6	0.083	0.885

Factor Score Coefficient Matrix

Variables	Factor 1	Factor 2
V1	0.358	0.011
V2	-0.001	0.375
V3	0.345	-0.043
V4	-0.017	0.377
V5	-0.350	-0.059
V6	0.052	0.395

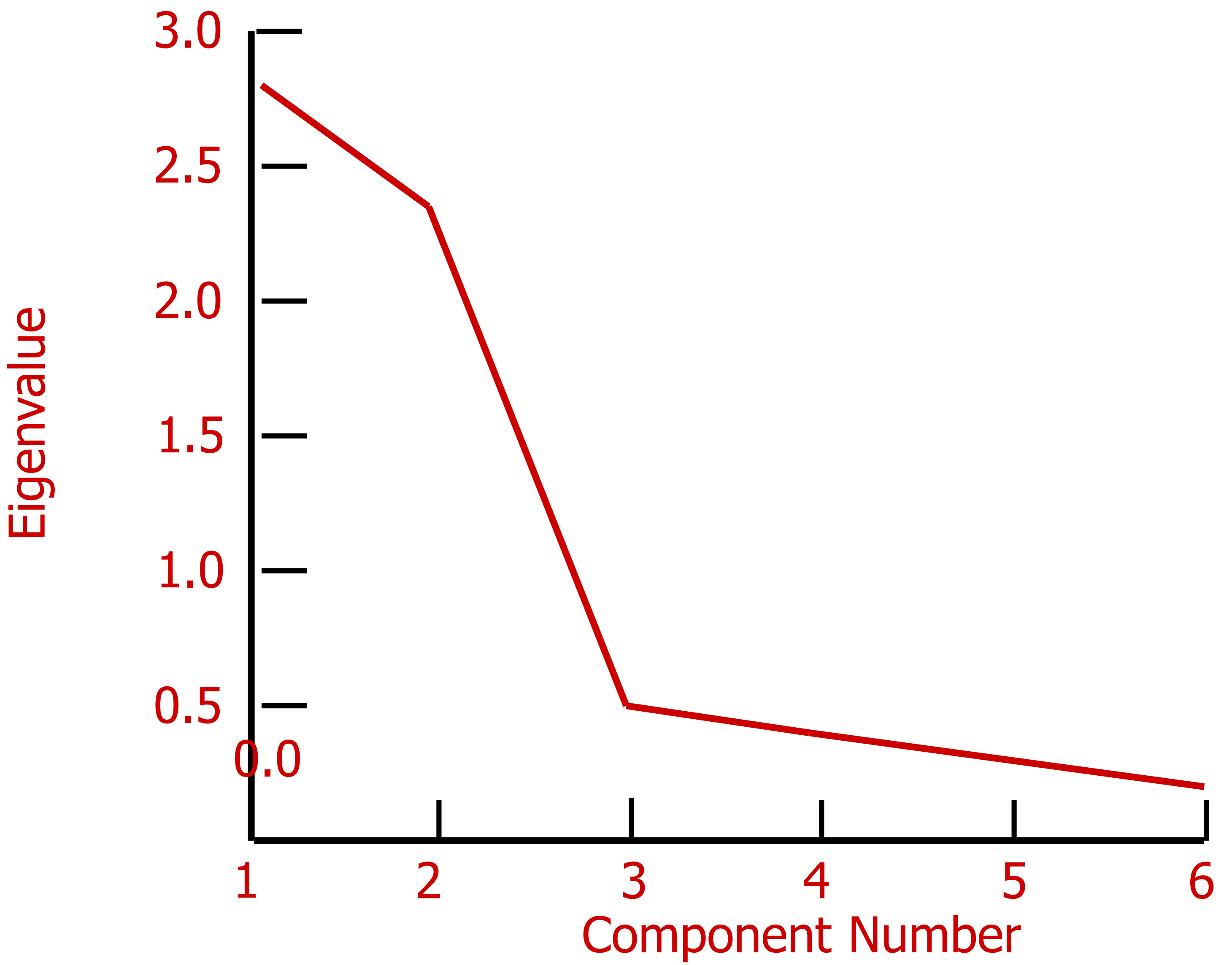
Results of Principal Components Analysis

-The lower-left triangle is correlation matrix;
-The diagonal has the communalities;
-The upper-right triangle has the residuals between the observed correlations and the reproduced correlations.

Factor Score Coefficient Matrix

Variables	V1	V2	V3	V4	V5	V6
V1	0.926	0.024	-0.029	0.031	0.038	-0.053
V2	-0.078	0.723	0.022	-0.158	0.038	-0.105
V3	0.902	-0.177	0.894	-0.031	0.081	0.033
V4	-0.117	0.730	-0.217	0.739	-0.027	-0.107
V5	-0.895	-0.018	-0.859	0.020	0.878	0.016
V6	0.057	0.746	-0.051	0.748	-0.152	0.790

Scree Plot



Factor Matrix Before and After Rotation

Factors

Variables	1	2
1	X	
2	X	X
3	X	
4	X	X
5	X	X
6		X

(a)

High Loadings
Before Rotation

Factors

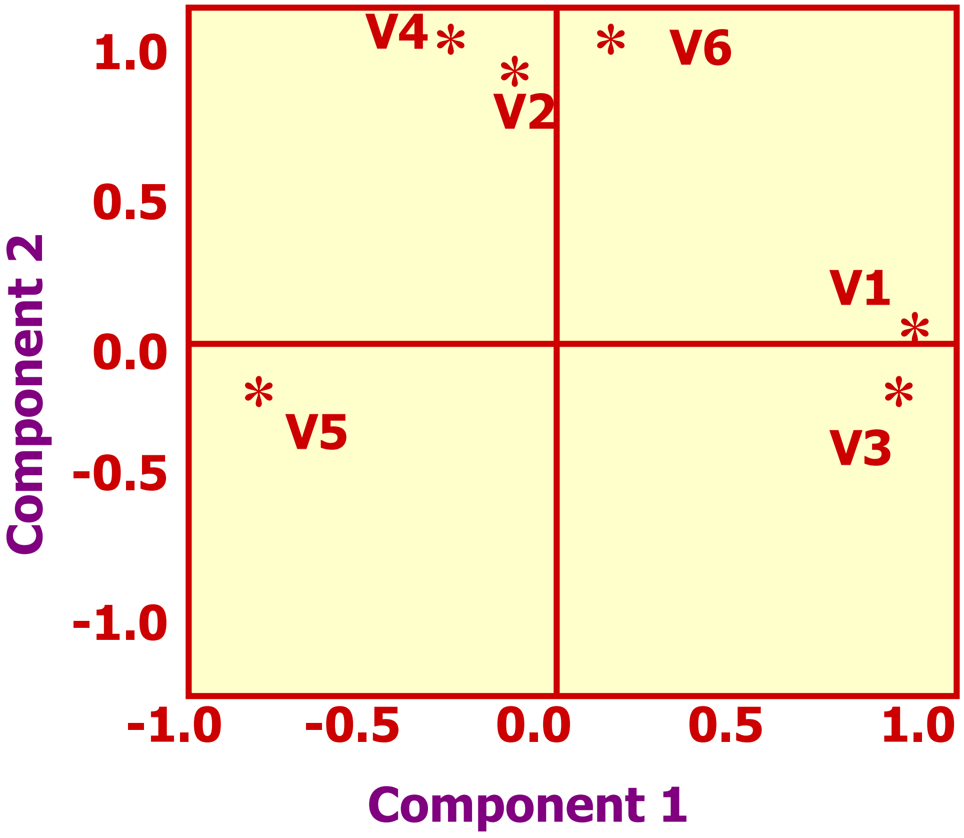
Variables	1	2
1	X	
2		X
3	X	
4		X
5	X	
6		X

(b)

High Loadings
After Rotation

Factor Loading Plot

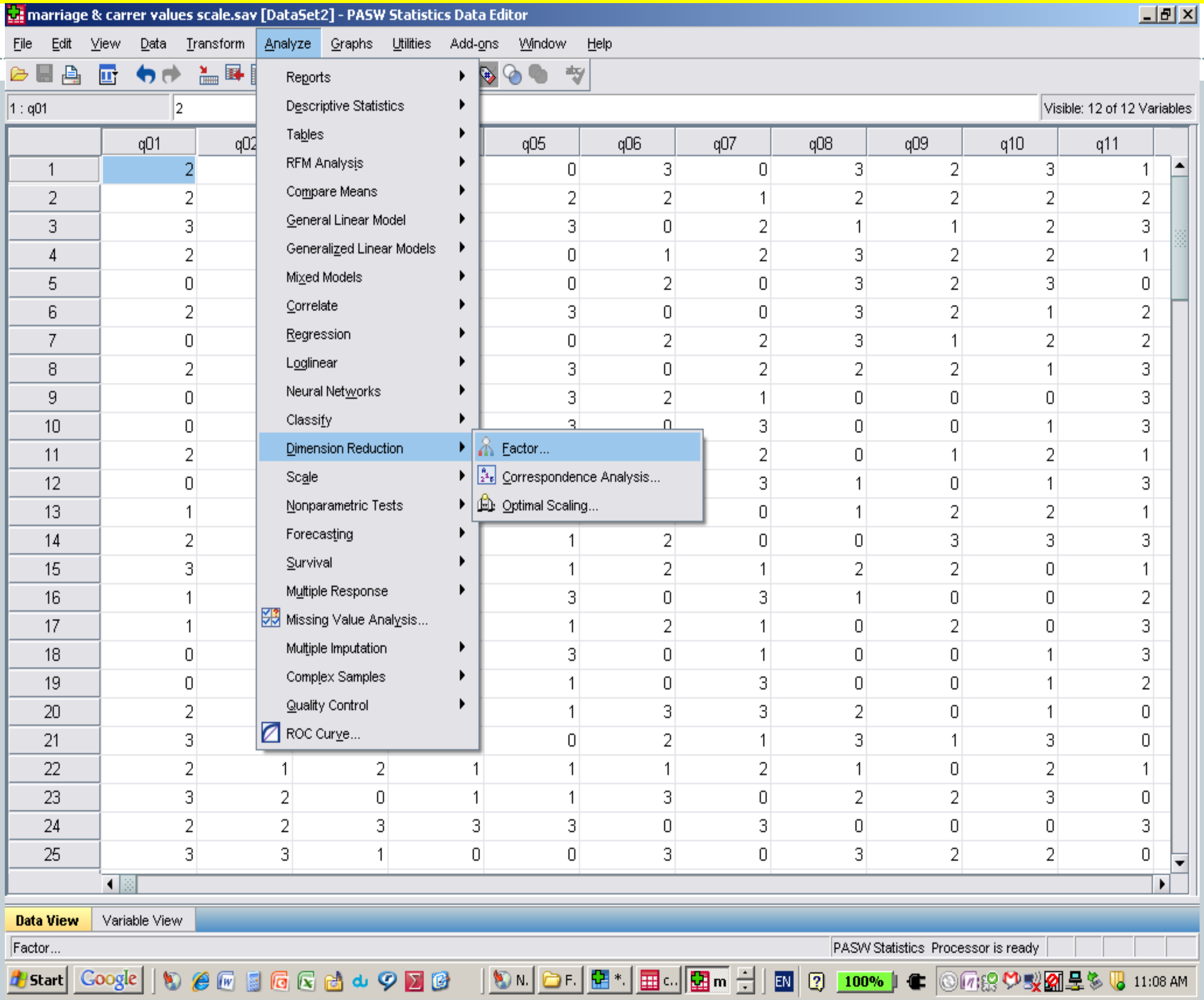
Component Plot in Rotated Space



Rotated Component Matrix

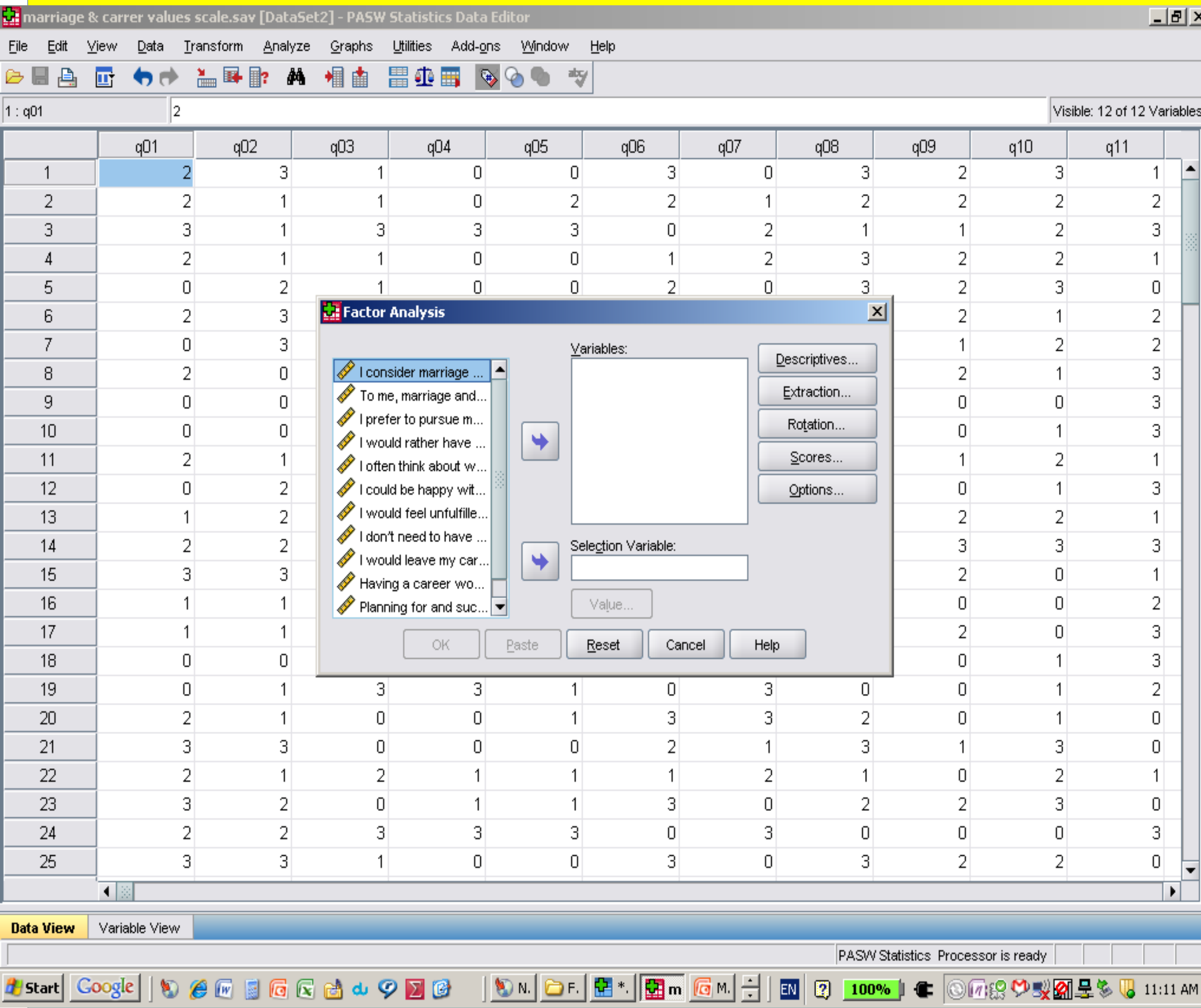
	Component	
	1	2
V1	0.962	-2.66E-02
V2	-5.72E-02	0.848
V3	0.934	-0.146
V4	-9.83E-02	0.854
V5	-0.933	-8.40E-02
V6	8.337E-02	0.885

Obtaining a Factor Analysis



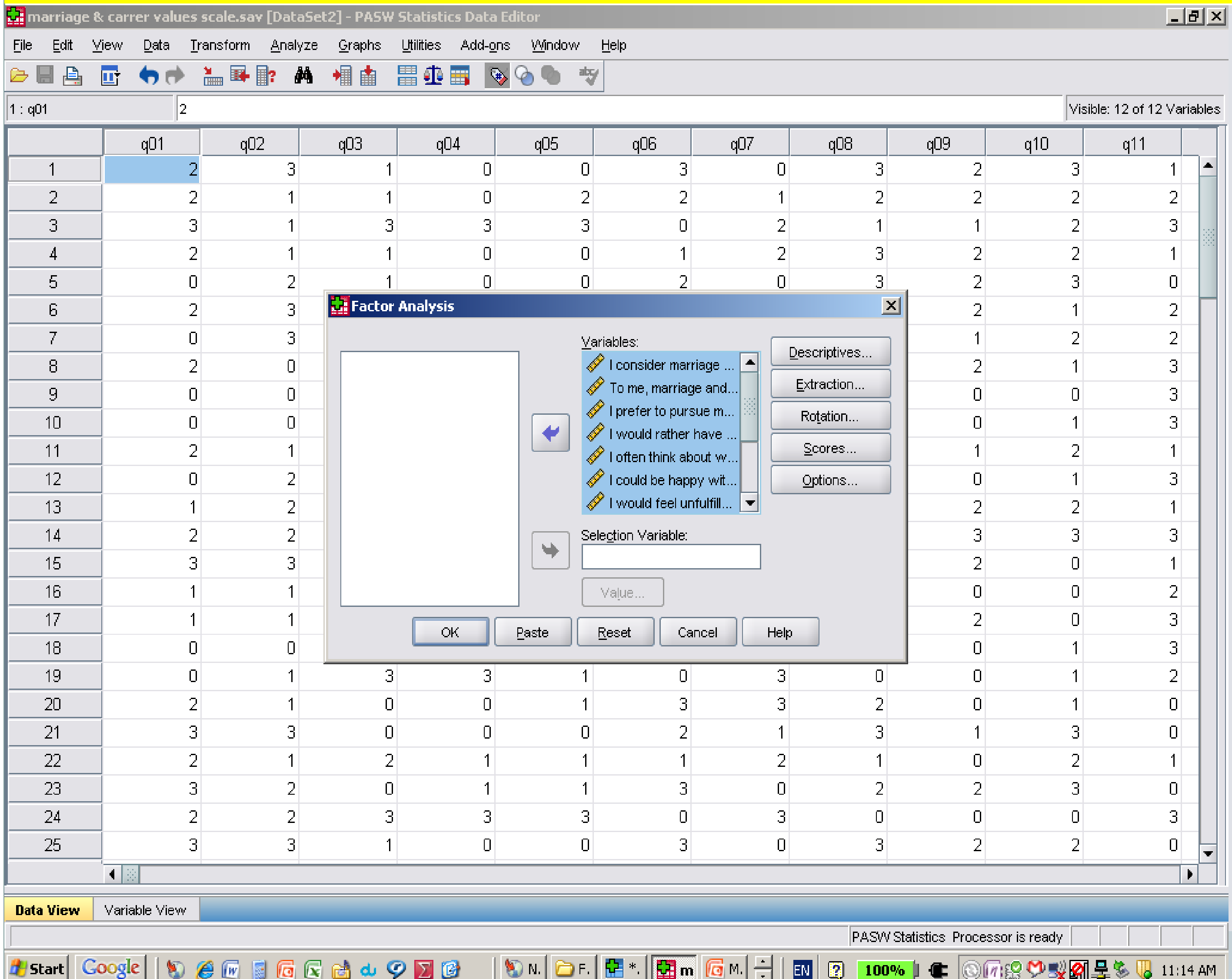
- **Click:**
  - **Analyze and select**
    - **Dimension Reduction**
    - **Factor**
    - **A factor Analysis Box will appear**

Obtaining a Factor Analysis



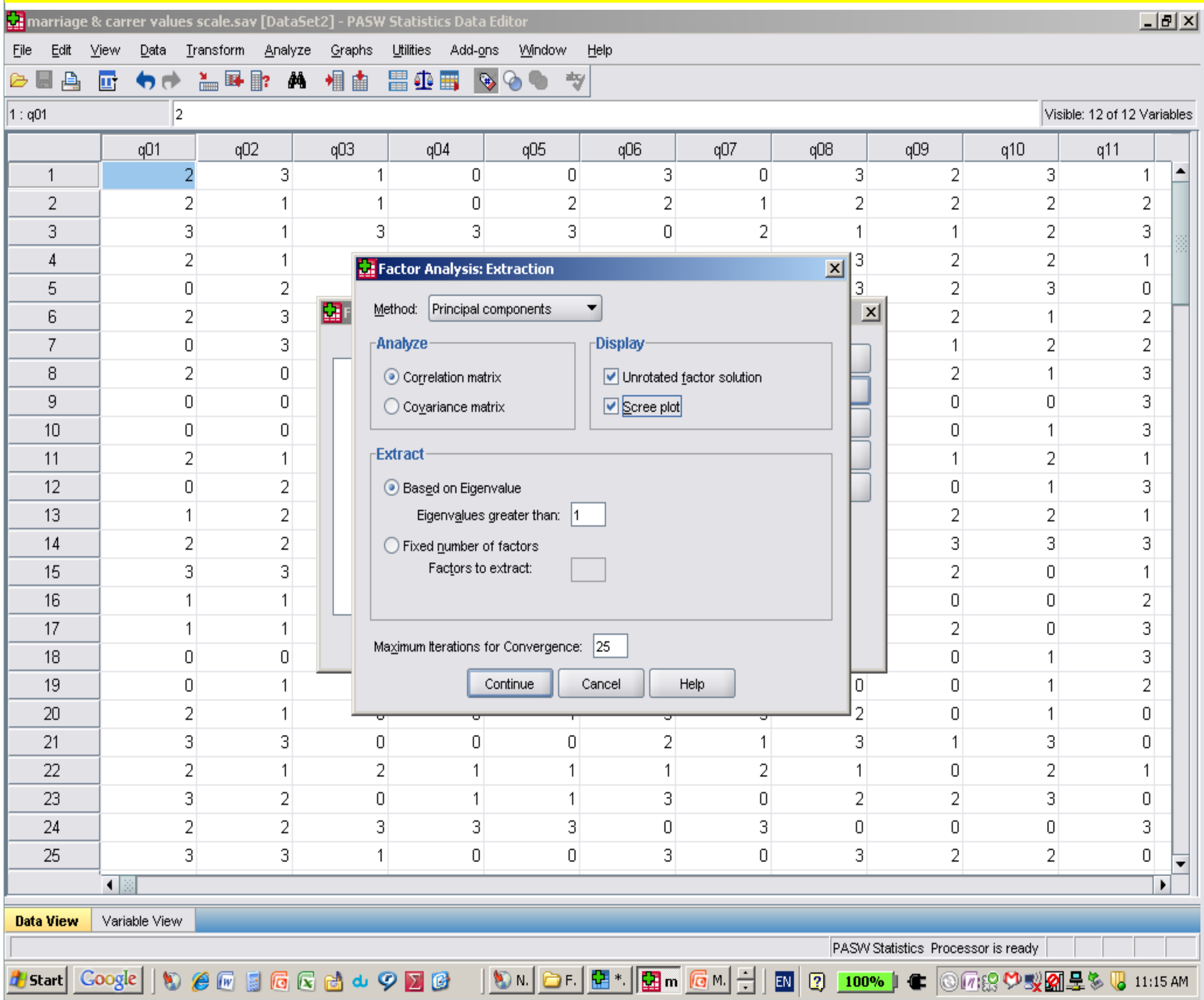
- Move variables/scale items to Variable box

Obtaining a Factor Analysis



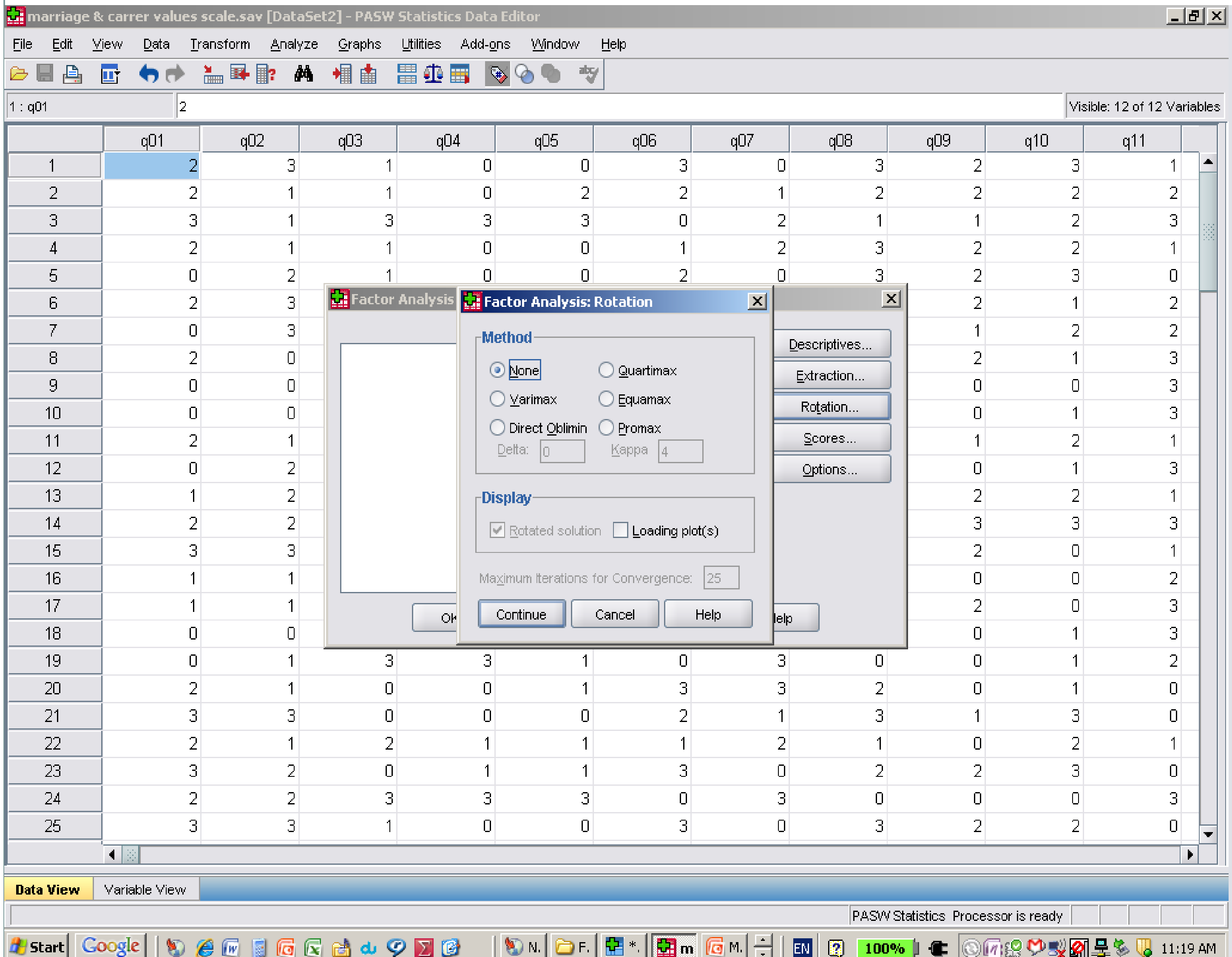
- Factor extraction
- When variables are in variable box, select:
  - Extraction

Obtaining a Factor Analysis



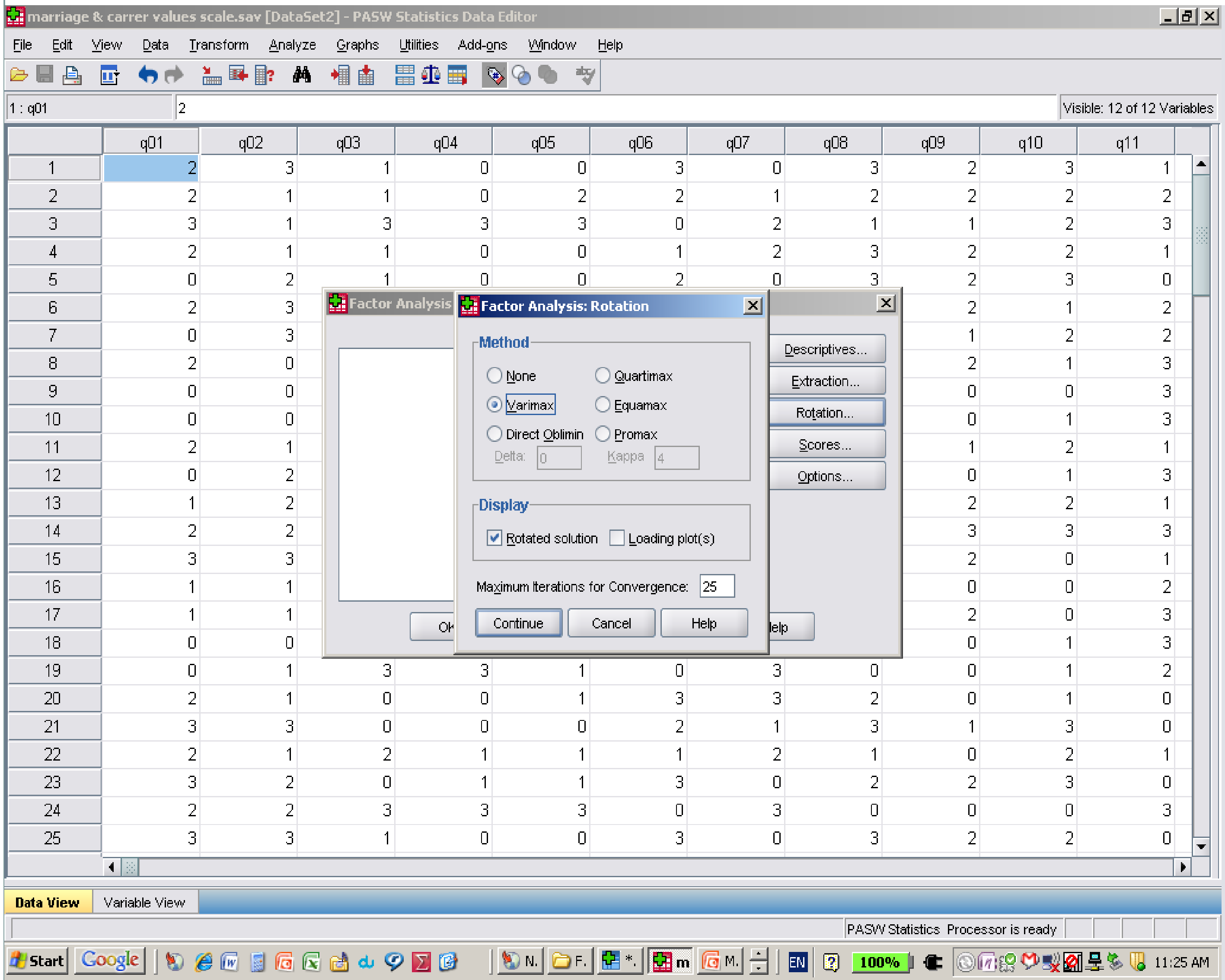
- When the factor extraction Box appears, select:
- Scree Plot
- keep all default selections including:
  - Principle component Analysis
  - Based on Eigen Value of 1, and
  - Un-rotated factor solution

Obtaining a Factor Analysis



- During factor extraction keep factor rotation default of:
  - None
  - Press continue

Obtaining a Factor Analysis



- During Factor Rotation:
- Decide on the number of factors based on actor extraction phase and enter the desired number of factors by choosing:
- Fixed number of factors and entering the desired number of factors to extract.
- Under Rotation Choose Varimax
- Press continue
- Then OK

Bibliographical References

- SPSS Base 7.0 Application Guide (1996). Chicago, IL: SPSS Inc.
- SPSS Base 7.5 For Windows User's Guide (1996). Chicago, IL: SPSS Inc.
- SPSS Base 8.0 Application Guide (1998). Chicago, IL: SPSS Inc.
- SPSS Base 8.0 Syntax Reference Guide (1998). Chicago, IL: SPSS Inc.
- SPSS Base 9.0 User's Guide (1999). Chicago, IL: SPSS Inc.
- SPSS Base 10.0 Application Guide (1999). Chicago, IL: SPSS Inc.
- SPSS Base 10.0 Application Guide (1999). Chicago, IL: SPSS Inc.
- SPSS Interactive graphics (1999). Chicago, IL: SPSS Inc.
- SPSS Regression Models 11.0 (2001). Chicago, IL: SPSS Inc.
- SPSS Advanced Models 11.5 (2002) Chicago, IL: SPSS Inc.
- SPSS Base 11.5 User's Guide (2002). Chicago, IL: SPSS Inc.
- SPSS Base 12.0 User's Guide (2003). Chicago, IL: SPSS Inc.
- SPSS 13.0 Base User's Guide (2004). Chicago, IL: SPSS Inc.
- SPSS Base 14.0 User's Guide (2005). Chicago, IL: SPSS Inc..
- SPSS Base 15.0 User's Guide (2007). Chicago, IL: SPSS Inc.
- SPSS Base 16.0 User's Guide (2007). Chicago, IL: SPSS Inc.
- SPSS Statistics Base 17.0 User's Guide (2007). Chicago, IL: SPSS Inc.
- Tabachnik, B.G., & Fidell, L.S. (2001). Using multivariate statistics (4^{th} Ed). Boston, MA: Allyn and Bacon.

QUESTIONS
OR
COMMENTS
?

Thank You