Eshrat Ara
Research Scholar,
Department of Psychology,
University of Kashmir,
Jammu and Kashmir, India

Exploring factors: Made easy

Eshrat Ara

Abstract
Factor analysis is used largely when the researcher has substantial numbers of variables seemingly measuring similar things. It has proven particularly useful with questionnaires. It examines the pattern of correlations between the variables and calculates new variables, which account for the correlations. In other words, it reduces data involving a number of variables down to a smaller number of factors which encompass the original variables. This paper presents a very brief overview on factor analysis. Although, the factor analysis is complicated and has many variations.

Keywords: Exploring factors, original variables, factor analysis

Introduction
Factor Analysis is not a new technique – it dates back to shortly after the First World War. It is an invention largely of psychologists, originally to serve a very specific purpose in the field of mental testing. It has proven generally more useful and is used in the development of psychological tests and questionnaires. Personality theories are heavily dependent on factor analysis (e.g., Raymond Cattell). The development of high-speed electronic computers has made the technique relatively routine since no longer does it requires months of hand calculations.

In the social sciences we are often trying to measure things that cannot be directly measured, they are so-called *latent variables*. For example 'personality', we can't measure personality directly: it has many facets. However, we can measure different aspects of personality: we could get some idea of motivation, stress levels, and so on. Having done this, it would be helpful to know whether these differences really do reflect a single variable. Put another way, are these different variables driven by the same underlying variable? So, factor analysis is a technique for identifying groups or clusters of variables.

This technique has three main uses:
1. To understand the structure of a set of variables. For example, pioneers of intelligence such as Spearman and Thurstone used factor analysis to try to understand the structure of the latent variable 'intelligence'.
2. To construct a questionnaire to measure an underlying variable. For example, we might design a questionnaire to measure burnout, personality, etc.
3. To reduce a data set to a more manageable size while retaining as much of the original information as possible. For example, multicollinearity can be a problem in multiple regression, and factor analysis can be used to solve this problem by combining variables that are collinear.

The logic and purpose of factor analysis is to process a large set of correlations revealing the dimensionality. In real data there are large numbers of correlations and it is impossible to analyze visually by simplistic eyeballing approach, and further the patterns of correlations are not clear in real data which complicates the process of dimensionality. There are three basic questions of Dimensionality: 1) number of dimensions, 2) correlation among dimensions (if more than one dimension is reflected), 3) psychological meaning of the dimensions. The answers to these questions require back-and-forth process.

Concepts in Factor Analysis
If we measure several variables, or ask someone several questions, the correlation between each pair of variables (or questions) can be arranged in what's known as an R-Matrix. An R-matrix is just a correlation matrix: a table of correlation coefficients between variables.

Correspondence
Eshrat Ara
Research Scholar,
Department of Psychology,
University of Kashmir,
Jammu and Kashmir, India

The existence of clusters of large correlation coefficients between subsets of variables suggests that those variables could be measuring aspects of the same underlying dimension. These underlying dimensions are known as factors or latent variables. In factor analysis we strive to reduce this R-matrix down to its underlying dimensions by looking at which variables seem to cluster together in a meaningful way. This data reduction is achieved by looking for variables that correlate highly with a group of other variables, but do not correlate with variables outside of that group.

An Example

R- Matrix																
	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16
V1	1.00	.39	.42	.31	.31	.19	.14	.23	.42	.30	.08	.22	.25	.18	.18	.19
V2	.39	1.00	.37	.30	.30	.23	.24	.26	.32	.20	.15	.12	.22	.25	.25	.14
V3	.42	.37	1.00	.43	.27	.28	.29	.21	.38	.34	.19	.31	.18	.33	.28	.29
V4	.31	.30	.43	1.00	.27	.30	.34	.14	.35	.32	.15	.19	.17	.32	.37	.28
V5	.31	.30	.27	.27	1.00	.41	.32	.23	.27	.33	.24	.22	.22	.32	.28	.25
V6	.19	.23	.28	.30	.41	1.00	.33	.21	.31	.22	.29	.27	.20	.27	.37	.29
V7	.14	.24	.29	.34	.32	.33	1.00	.23	.22	.24	.44	.17	.30	.35	.41	.24
V8	.23	.26	.21	.14	.23	.21	.23	1.00	.16	.22	.21	.19	.25	.25	.14	.14
V9	.42	.32	.38	.35	.27	.31	.22	.16	1.00	.24	.17	.25	.26	.26	.23	.27
V10	.30	.20	.34	.32	.33	.22	.24	.22	.24	1.00	.24	.31	.15	.24	.24	.32
V11	.08	.15	.19	.15	.24	.29	.44	.21	.17	.24	1.00	.24	.26	.20	.41	.11
V12	.22	.12	.31	.19	.22	.27	.17	.19	.25	.31	.24	1.00	.19	.31	.12	.26
V13	.25	.22	.18	.17	.22	.20	.30	.25	.26	.15	.26	.19	1.00	.33	.20	.17
V14	.18	.25	.33	.32	.32	.27	.35	.25	.26	.24	.20	.31	.33	1.00	.24	.39
V15	.18	.25	.28	.37	.28	.37	.41	.14	.23	.24	.41	.12	.20	.24	1.00	.26
V16	.19	.14	.29	.28	.25	.29	.24	.14	.27	.32	.11	.26	.17	.39	.26	1.00

It is not so simple. Real data usually include many more variables. Large number of variables produce, a much larger number of correlations to examine. And then ambiguity complicates the process of evaluation. Factor analysis is a statistical procedure that simplifies this process. Factor analysis is a family of procedures like Exploratory Factor Analysis, Principle Component Analysis, Confirmatory Factor Analysis, etc.

The Questions Addressed By Factor Analysis
Q1: The first question addressed by the factor analysis is the number of factors. To address this question, we often refer to the statistics called Eigenvalues. An eigenvalue is a measure of variance explained in the vector space created by the factors. Eigenvalues indicate the substantive importance of that factor. We retain only factors with larger eigenvalues but to decide eigenvalue is large is complex. So a second common way of using eigenvalues is Scree Plot, which is a graphical presentation of eigenvalues. Cattell technique is to plot a graph of each eigenvalue (Y-axis) against the factor associated (X-axis). Through Scree plot the relative importance of each factor becomes apparent. Cattell suggested that the cut-off point for selecting factors be the point of inflexion of the curve, which is where the slop of the line changes dramatically but do not include factor at the point of inflexion itself, as the point of inflexion represents an error factor, only factors to the left are extracted. With sample of more than 200 participants, Scree plot gives fairly reliable criterion for factor selection.

An Example

Total Variance Explained			
Components	**Initial Eigenvalues**		
	Total	**%**	**Cumulative %**
1	4.917	30.733	30.733
2	1.363	8.519	39.252
3	1.095	6.844	46.096
4	1.062	6.641	52.737
5	.903	5.646	58.383
6	.836	5.225	63.608
7	.812	5.075	68.683
8	.723	4.517	73.200
9	.682	4.261	77.461
10	.618	3.864	81.325
11	.579	3.617	84.942
12	.545	3.406	88.348
13	.515	3.219	91.567
14	.487	3.045	94.611
15	.440	2.749	97.360
16	.422	2.640	100.000

An Example
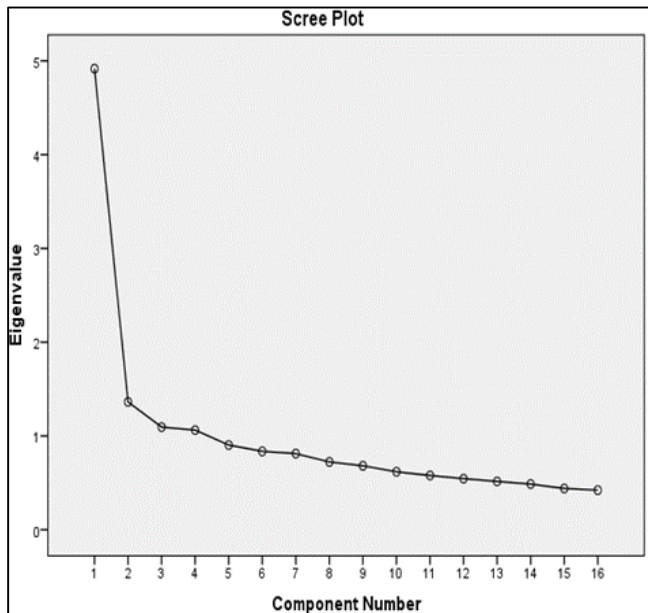
Kaiser (1960) recommended retaining all factors with eigenvalue greater than 1 because eigenvalues represent the amount of variation explained by a factor and an eigenvalue of 1 represents a substantial amount of variation. According to Thurstone, it is better to retain too few rather than too many factors. Jolliffe (1986) says that Kaiser's criterion is too strict and suggests the third option of retaining all factors with Eigenvalue more than 0.7, with variables less than 30 and communality after extraction all greater than 0.7, as evidence of accuracy. Kaiser's criterion is also accurate when sample size exceeds 250 and average communality is greater than or equal to 0.6. It is best to use Scree plot when sample size exceeds 200.

With 30 or more variables and communalities greater than 0.7 for all variables, different solutions are unlikely. Fewer than 20 variables and any low communalities (< 0.4) differences can occur. Communality refers to common variance, which is first assumed to be 1. Communality after extraction will always be less than 1. Better to be close to 1. When there are more factors, communality will be greater because less information is discarded. The final decision depends on why are we doing analysis, if we are trying to overcome multicollinearity problem in regression then, better to extract too many.

Q2: The second question concerns the correlation among factors. This question is addressed through a process called Factor Rotation. Once factors are extracted, we can calculate to what degree variables load onto these factors. Factor Rotation is used to discriminate between factors, such that variables are loaded maximally to one factor. We have two types of rotation:

1. Orthogonal rotation simply means that none of the factors are actually allowed to correlate with each other. In other words, orthogonal means unrelated, that is factors are rotated while keeping them independent or unrelated. Before rotation, all factors are independent; they do not correlate at all and orthogonal rotation ensures that the factors remain uncorrelated. So axes are turned while remaining perpendicular.

2. Oblique rotation means that the factors are allowed to correlate with each other (although they can end up uncorrelated). SPSS has three methods of orthogonal rotation (varimax, quartimax and equamax) and two methods of oblique rotation (direct oblimin and promax). The resulting output depends on which method we select.

The choice of rotation depends on theoretical reason and on how variables cluster on the factors before rotation. One approach is to run the both, if oblique rotation shows negligible correlation then it is reasonable to use orthogonal rotation solution. If there is sound theoretical reason then use oblique rotation. In oblique rotation Factor Transformation Matrix is needed.

For naturalistic data and certainly data involving humans, some argue that orthogonal rotation should never be used because one cannot think of any psychological construct that is not in any way correlated with some other psychological construct.

Further, Unrotated Factor Solution is useful in assessing the improvement of interpretation due to rotation. If rotated solution is little better than unrotated solution then it indicates inappropriate rotation method used.

An Example
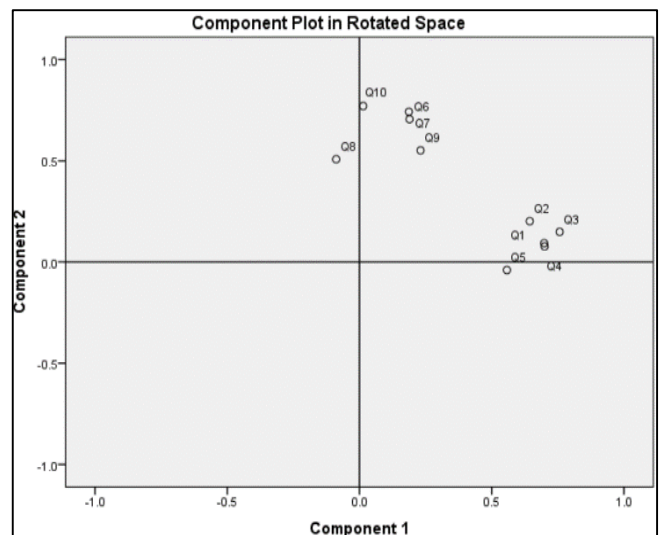
Q3: The third question concerns the psychological meaning of the factors. This is addressed through the analysis of values called Factor Loadings. Factor Loadings represent the link between variables and factors, based on which we interpret the psychological meaning of each factor. It indicates the importance of a variable to a factor which is assessed through statistical significance of a factor loading which is correlation or regression coefficient. Factor Structure Matrix presents the correlations between variables and each factor. There are other forms of factor loadings also.

Typically, loading of an absolute value of more than 0.3 is taken to be important. But it depends on the sample size also. Stevens (2002) recommends, that for a sample size of 50, a loading of 0.72 can be considered significant; for 100, it should be greater than 0.51, for 200; greater than 0.36; for 300, greater than 0.29; for 600, greater than 0.21; for 1000, greater than 0.16. These values are based on an alpha level 0.1 (two-tailed). Therefore large samples, small loadings are

statistically meaningful. But the significance of a loading gives little indication of the importance of a variable to a factor. So, Stevens (2002) recommends interpreting only factor loadings with an absolute value greater than 0.4 which explains around 16% of the variance in the variable.

An Example

Factor Structure Matrix		
	Component	
	1	2
V1	.773	.118
V2	.711	.310
V3	.680	.332
V4	.609	.474
V5	.443	.472
V6	.343	.588
V7	.265	.757
V8	.303	.226
V9	.653	.278
V10	.404	.310
V11	.077	.717
V12	.199	.163
V13	.280	.343
V14	.307	.385
V15	.340	.775
V16	.271	.300
Extraction Method: Principal Component Analysis. Rotation Method: Oblimin with Kaiser Normalization.		

An Example

Component Plot in Rotated Space
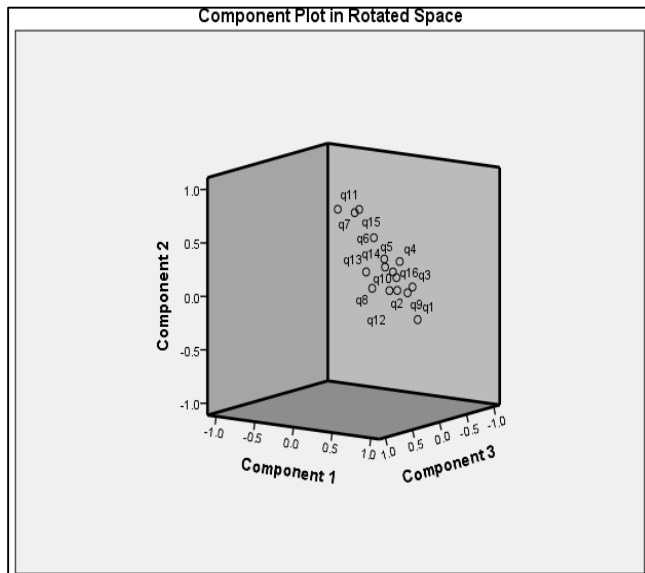
Stages and Issues in Factor Analysis
1. The first stage of factor analysis is to scan the variables to check that they relate to each other to some degree but not too strongly. It concerns with Reliability Analysis. Too high correlations and too low correlations are two problems. The remedy is to remove them from the analysis. But it is subjective. It depends on the data set also.
   Too high correlations indicate Multicollinearity. If it is mild, it is not a problem for PCA, but should be avoided. When there is perfect correlation between variables, it indicates Singularity.

If variables in an R-matrix did not correlate at all, then correlation matrix would be an Identity Matrix, i.e., the off-diagonal components are (0) zero. Bartlett's test examines whether the R-Matrix resembles identity matrix (independent variables). When there are no correlations, there cannot be clusters, so no factors. If Bartlett's test is significant, it means that correlations between variables are overall significantly different from zero, which is good. There are significance tests for Multicollinearity like Haitovsky test. One rule is to check the Determinant of the R-matrix, which should be greater than 0.00001.

2. The factor analysis itself has several stages: Check some initial issues
   ✓ To perform a factor analysis, all of the variables in the analysis must be measured continuously, (i.e., intervally scaled).
   ✓ Variables in the analysis must be having normal distributions.
   ✓ It is also important to have a large enough sample size. Correlation coefficients fluctuate from sample to sample, more in small sample size than in large sample size. Common rule is to suggest 10 to 15 participants per variable. Kass and Tinseley (1979) suggests 5 to 10 participants per variable, upto a total of 300 beyond which test parameters tend to be stable regardless of the participant to variable ratio. Tabachnick and Fidell (2007) agree at least 300 sample size for factor analysis, which is good. With all communalities 0.6, small size may be adequate. Another alternative is to use the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy. KMO can be calculated for individual and multiple variables and represents the ratio of the squared correlation between variables to the squared partial correlation between variables. It varies from 0-1. 0 indicates sum of partial correlations is large relative to the sum of correlations, indicating diffusion and so factor analysis is inappropriate. Values close to 1 indicates that patterns of correlations are compact and so distinct and reliable factors will be yielded. Kaiser recommends value > 0.5 acceptable, 0.5 to 0.7 mediocre, 0.7 to 0.8 great, and value > 0.9 superb.
3. Decide how many factors to retain.
4. Finally decide which items load onto which factors.
5. Try to make sense of the meaning of the factors. It concerns the nature of the dimensions, i.e, what do the dimensions represent. It requires in-depth examination. It concerns with Test validity. However factor analysis provides information that is potentially important part of this examination. Values called factor loadings convey this information, based on which we begin to interpret the psychological meaning of each dimension.
6. Having done all that you can consider whether the items you have are reliable measures of what you're trying to measure.
   Factor Analysis can be performed by many statistical packages e.g., SPSS, AMOS, SAS, MPLUS, MINITAB, LISERAL, etc.

Reference
1. Field A. Discovering Statistics Using SPSS (Third Ed.). Sage Publications LTD, 2009.
2. Urdan TC. Statistics in Plain English (Third Ed.).

Routledge Taylor and Francis Group, 2010.
3. Howitt D, Cramer D. Introduction to Statistics in Psychology (Fifth Ed.). Pearson Education Limited, 2011.
4. Furr RM, Bacharach VR. Psychometrics: An Introduction. SAGE Publications, 2008.