UNIVERSITAT ROVIRA I VIRGILI

How to report the percentage of explained common variance in exploratory factor analysis

Urbano Lorenzo-Seva

Tarragona 2013

Please reference this document as:

Lorenzo-Seva, U. (2013). *How to report the percentage of explained common variance in exploratory factor analysis. Technical Report*. Department of Psychology, Universitat Rovira i Virgili, Tarragona.

Document available at: http://psico.fcep.urv.cat/utilitats/factor/

Contents

1. Percentage of explained variance as an index of goodness of fit

A popular and intuitive index of goodness of fit in multivariate data analysis is the percentage of explained variance: the higher the percentage of variance a proposed model manages to explain, the more valid the model seems to be. In this document we study how this index can be reported in the context of exploratory factor analysis.

Table 1. Univariate descriptive statistics

Variable	Mean	Standard deviation	Variances
1	49.95	3.26	10.63
2	60.06	5.02	25.20
3	45.01	2.20	4.84
4	98.57	10.96	120.12
5	100.05	10.95	119.90
6	49.45	7.06	49.84
7	30.50	4.01	16.08
8	44.56	5.62	31.58

To make the text more compressive, we base our explanations on the analysis of a particular dataset of eight observed variables: the mean, standard deviation, and variances are shown in Table 1. We suspect that an unknown number of latent variables may explain the relationship between the eight observed variables. In multivariate data analysis the relationship between observed variables is typically described using the standardized variance/covariance matrix (i.e., the correlation matrix shown in Table 2). As can be observed the value of the variances in the correlation matrix is 1 for all the variables (and not the variance values shown in Table 1): the reason for this is that the variables have been standardized. As this is quite a *simple* dataset, the mere visual inspection of the correlation matrix provides important information:

1. Variables 1, 2, 3, and 4 seem to be mainly related among themselves, as are variables 5, 6, 7. These two independent clusters of variables could be explained by the existence of two independent latent variables, each of which is responsible for the variability of one cluster of observed variables.

2. Even if two clusters of observed variables seem to exist in the data, the correlation values among variables are systematically low. This result indicates that the observed variables in each cluster do not share a large amount of variance (i.e., the amount of common variance, also known as *communality*, is low).

Table 2. Correlation matrix among the eight variables. Correlation values larger than .20 are printed in bold

	V1	V2	V3	V4	V5	V6	V7	V8
V1	1	**.3683**	.1918	**.2746**	.0852	.0844	.0223	.2096
V2	**.3683**	1	.1746	.1103	.1646	.1806	.0761	**.2403**
V3	.1918	.1746	1	**.2105**	.0520	-.0147	.0273	.1466
V4	**.2746**	.1103	**.2105**	1	.0868	-.0008	-.0034	.0759
V5	.0852	.1646	.0520	.0868	1	.1793	**.4105**	**.4761**
V6	.0844	.1806	-.0147	-.0008	.1793	1	.1692	**.2203**
V7	.0223	.0761	.0273	-.0034	**.4105**	.1692	1	**.3873**
V8	.2096	**.2403**	.1466	.0759	**.4761**	**.2203**	**.3873**	1

If the aim is to make an analytical study of the information in the correlation matrix in terms of the underlying latent variables, the most appropriate technique available in multivariate data analysis is Exploratory Factor Analysis (EFA). The aim of EFA is to determine the latent structure of a particular dataset by discovering common factors (i.e., the latent variables). In this regard, EFA accounts for the common variance (i.e., the shared variance among observed variables). In the analysis, the common variance is partitioned from its unique variance and error variance, so that only the common variance is present in the factor structure: this means that the percentage of explained variance should be reported in terms of common variance (i.e., the percentage of explained common variance should be reported).

Researchers often compute Principal Component Analysis (PCA) as an approximation of EFA. The aim of PCA is to explain as much of the variance of the observed variables as possible using few composite variables (usually referred to as components). PCA does not discriminate between common variance and unique variance. Whether PCA is a proper approximation of EFA or not is a controversy on which *Multivariate Behavioral Research* published a special issue edited by Dr. Mulaik (1992). Thompson (1992) argued that the practical difference between the approaches (PCA vs EFA) is often negligible in terms of interpretation. On the other hand, Gorsuch (1986)

concluded that the differences in results decrease as (a) the score reliability of the measured variables increases, or (b) the number of variables measured increases. Snook and Gorsuch (1989) added that, when only a few variables are being analyzed or the communality is low, PCA and EFA analytic procedures produce divergent results.

In the problem that concerns us (reporting the percentage of explained variance), computing PCA is appealing because: (a) the percentage of explained variance is an immediate index of goodness of fit in PCA; and (2) it is not obvious how to compute the percentage of explained common variance in EFA. Unfortunately, our dataset encounters situations (few observed variables and low communality) in which PCA is not an appropriate approach to EFA.

Given our pedagogical aim, the rest of this document focuses on: (a) how to obtain the percentage of explained variance in PCA; (b) why it is not possible to compute the percentage of explained common variance in most factor methods; (c) how to compute the percentage of explained common variance in an EFA; and (d) the advantages of being able to report the percentage of explained common variance in an EFA.

2. Percentage of explained variance in principal component analysis

PCA aims to summarise the information in a correlation matrix. The total amount of variance in the correlation matrix can be calculated by adding the values on the diagonal: as each element on the diagonal has a value of 1, the total amount of variance also corresponds to the number of observed variables. In our dataset, the total amount of variance is 8. This total amount of variance can be partitioned into different parts where each part represents the variance of each component. The eigenvalues printed in Table 3 represent the amount of variance associated with each component. If the eigenavalues are added, the resulting total should be the total variance in the correlation matrix (i.e., the addition 2.244 + 1.4585 + ... + 0.4866 should be equal to 8). The percentage of explained variance of each component can be easily computed as the corresponding eigenvalue divided by the total variance: for example, the percentage of variance explained by the first component is 2.224 / 8 = .28 (or in terms of percentage 28%). The first component also counts for 28% of the variance. When the percentage of explained variance is reported for a particular dataset, the value that is actually reported is the addition of the percentages of the explained variance for each of the components retained (i.e., the accumulated percentage of explained variance). Table 3 shows that if the aim were to explain 100% of the variance in the correlation matrix, then we would need to retain as

many components as observed variables (which would make no sense at all). However, the idea is to select an *optimal* number of components. The optimal number of components can be defined as the minimum number of components that accounts for the maximum possible variance.

In our visual inspection of the correlation matrix in Table 2, we already intuited that retaining two components should be enough for our dataset. However, if only two components are retained the (accumulated) percentage of explained variance (46.3%) would suggest a poor fit of the component solution. To achieve an acceptable fit, it seems that we should retain at least five components. It seems clear that the percentage of explained variance does not suggest an optimal number of components to be retained.

Table 3. Eigenvalues and percentages of variance associated with each component

Component	Eigenvalue	Percentage of explained variance	Accumulated percentage of explained variance
1	2.2440	28.0	28.0
2	1.4585	18.2	46.3
3	0.9996	12.5	58.8
4	0.8232	10.3	69.1
5	0.7933	9.9	79.0
6	0.6064	7.6	86.6
7	0.5883	7.4	93.9
8	0.4866	6.1	100.0

The reason why the percentage of explained variance does not properly describe the goodness of fit is because PCA is not a proper approximation of EFA for this dataset. If we continue the analysis with our initial idea of retaining two components and we rotate the loading matrix with varimax (Kaiser 1958), the simplicity of the component solution that we obtain seems to reinforce our initial intuition (i.e., of retaining only two components). Table 4 shows this rotated two-component solution.

Table 4. Loading matrix of component solution after Varimax rotation. Salient loading values are printed in bold

Variables	Component 1	Component 2
v1	**.75**	.09
v2	**.60**	.26
v3	**.58**	.00
v4	**.62**	-.05
v5	.06	**.77**
v6	.05	**.46**
v7	-.10	**.74**
v8	.23	**.75**

In a typical real situation, probably involving many more latent variables, few observed variables per latent variable, and low communality, the visual inspection of the correlation matrix would be useless. In addition, as we have seen in our example, PCA would not help us to take the proper decisions either. In a situation such as this, the wisest decision would be to compute the most appropriate technique available in multivariate data analysis when the aim is to study the information in the correlation matrix in terms of the underlying latent variables (i.e., to compute an EFA).

3. Percentage of explained common variance in exploratory factor analysis

As mentioned above, in EFA only the common variance is present in the factor structure, and the percentage of explained variance should be reported in terms of common variance (i.e., the percentage of explained common variance). However, the percentage of explained common variance cannot be computed in most factor analysis methods. To show this, we analyze our dataset using Principal Axis Factor (PAF) analysis in the section below. We decided to use PAF because it is quite a straightforward method, but the conclusion that we draw can be generalized to most factor analysis methods (like Unweighted Least Squares factor analysis, or Maximum Likelihood factor analysis). The only method that enables the percentage of explained common variance to be comuted is Minimum Rank Factor Analysis (MRFA).

3.1. Principal axis factor analysis

As mentioned above, PCA analyzes the variance contained in a correlation matrix. In EFA, the matrix that is analyzed is known as the reduced correlation matrix: the diagonal elements of the correlation matrix are substituted by estimates of the communality of each variable. In PAF, the multiple correlation value is typically used as an estimate of communality. Table 5 shows the reduced correlation matrix with the multiple correlation values already placed on the diagonal of the matrix.

Table 5. Reduced correlation matrix analyzed in principal axis factor analysis. Multiple correlation values are printed in bold

	V1	V2	V3	V4	V5	V6	V7	V8
V1	**.2128**	.3683	.1918	.2746	.0852	.0844	.0223	.2096
V2	.3683	**.1898**	.1746	.1103	.1646	.1806	.0761	.2403
V3	.1918	.1746	**.0883**	.2105	.0520	-.0147	.0273	.1466
V4	.2746	.1103	.2105	**.1073**	.0868	-.0008	-.0034	.0759
V5	.0852	.1646	.0520	.0868	**.2970**	.1793	.4105	.4761
V6	.0844	.1806	-.0147	-.0008	.1793	**.0820**	.1692	.2203
V7	.0223	.0761	.0273	-.0034	.4105	.1692	**.2251**	.3873
V8	.2096	.2403	.1466	.0759	.4761	.2203	.3873	**.3273**

In a correlation matrix, the total amount of variance is obtained by adding the values on the diagonal of the matrix. In a reduced correlation matrix, the total amount of variance is obtained in the same way. The total amount of variance of the reduced correlation matrix shown in Table 5 is 1.5296 (i.e., the addition of the multiple correlation values .2128 + .1898 + ... + .3273). It should be noted that this is the total amount of common variance. As the total amount of common variance can be readily obtained, the strategy used in PCA to obtain percentages of explained variance could be replicated: (a) to compute the eigenvalues, and (b) to use them as partitions of the common variance. The eigenvalues related to the reduced correlation matrix are shown in Table 6. By adding the eigenvalues, we can confirm that they do indeed add up to the total amount of common variance (i.e., the addition 1.4862 + .6434 + ... - .2676 equals 1.5296). However, there is an important limitation: some of the eigenvalues are negative (i.e., the reduced correlation matrix is said to be non-positive definite). This means that these eigenvalues

cannot be safely interpreted as partitions of the common variance, and the percentages of explained common variances cannot be computed.

Table 6. Eigenvalues associated with the reduced correlation matrix

Factor	Eigenvalue
1	1.4862
2	0.6434
3	0.1322
4	-0.0599
5	-0.0916
6	-0.1488
7	-0.1643
8	-0.2676

The reduced correlation matrix computed in most factor analysis methods is systematically non-positive definite. The typical conclusion is that the percentage of explained common variance cannot be computed in EFA. This is why computer software packages refuse to compute the explained common variance as an index of goodness of fit, and do not even include the eigenvalues associated with the reduced correlation matrix in the output.

3.2. Minimum rank factor analysis

Minimum Rank Factor Analysis (MRFA, Ten Berge & Kiers, 1991) also analyzes a reduced correlation matrix. However, in MRFA the estimates of the communality to be used on the diagonal of the reduced correlation matrix are carefully chosen so that the reduced correlation matrix is positive definite (i.e., none of the related eigenvalues is negative): the greatest lower bound to reliability (Woodhouse & Jackson, 1977; Ten Berge, Snijders & Zegers,1981) is used as an estimate. Table 7 shows the reduced correlation matrix with the estimates of communality already on the diagonal of the matrix. The total amount of common variance of the reduced correlation matrix shown in Table 7 is 3.4413 (i.e., the addition of values .5272 + .5007 + ... + .5393).

Table 7. Reduced correlation matrix analyzed in minimum rank factor analysis. Estimated communality values are printed in bold

	V1	V2	V3	V4	V5	V6	V7	V8
V1	**.5272**	.3683	.1918	.2746	.0852	.0844	.0223	.2096
V2	.3683	**.5007**	.1746	.1103	.1646	.1806	.0761	.2403
V3	.1918	.1746	**.3110**	.2105	.0520	-.0147	.0273	.1466
V4	.2746	.1103	.2105	**.3566**	.0868	-.0008	-.0034	.0759
V5	.0852	.1646	.0520	.0868	**.6527**	.1793	.4105	.4761
V6	.0844	.1806	-.0147	-.0008	.1793	**.2245**	.1692	.2203
V7	.0223	.0761	.0273	-.0034	.4105	.1692	**.3293**	.3873
V8	.2096	.2403	.1466	.0759	.4761	.2203	.3873	**.5393**

The eigenvalues associated with the reduced correlation matrix are shown in Table 8. As can be observed, none of the eigenvalues is negative. Furthermore, the total of all the eigenvalues is equivalent to the total amount of common variance (i.e., the addition 1.7385 + 0.9063 + ... + 0 equals 3.4413). In conclusion, then, the eigenvalues can be considered as partitions of the total common variance, and the percentage of explained common variance can be easily computed as the corresponding eigenvalue divided by the total common variance: for example, the percentage of common variance explained by the first factor is 1.7385 / 3.4413 = .505 (or in terms of percentage 50.5%).

Table 8. Eigenvalues and percentages of explained common variance associated with each factor

Factor	Eigenvalue	Percentage of explained common variance	Accumulated percentage of explained common variance
1	1.7385	50.5	50.5
2	0.9063	26.3	76.9
3	0.3764	10.9	87.8
4	0.1859	5.4	93.2
5	0.1289	3.7	96.9
6	0.1052	3.1	100.0
7	0.0000	0.0	100.0
8	0.0000	0.0	100.0

Table 8 shows that the extraction of two factors accounts for 76.9% of the common variance: this means that a two-factor model is associated with a percentage of explained common variance of 76.9%. As can be seen, using the percentage of common variance as a goodness of fit index helps to correctly assess the most suitable factor model for our dataset.

Let us continue the analysis: we extract two factors and rotate the loading matrix with Varimax (Kaiser 1958). Table 9 shows the rotated two-factor solution. Again, the simplicity of the factor solution seems to reinforce our initial intuition (i.e., to retain only two factors).

Table 9. Loading matrix of factor solution after Varimax rotation. Salient loading values are printed in bold

Variables	Factor 1	Factor 2
v1	**.67**	.08
v2	**.54**	.22
v3	**.39**	.05
v4	**.44**	.00
v5	.04	**.76**
v6	.09	**.31**
v7	-.04	**.56**
v8	.22	**.67**

If we compare the two-component solution in Table 4, and the two-factor solution in Table 9, we realize that the salient loading values associated with the component solution are larger than the salient loading values associated with the factor solution. In addition, the Factor Simplicity Index (Lorenzo-Seva, 2003) shows that the two-component solution has the simplest structure (a value of .6342 for the component solution versus a value of .5951 for the factor solution). However, this apparently better component solution is artificial: as PCA confounds common variance and unique variance the loading values appear larger than they should actually be. For example, Widaman (1993) showed that when the data fit the assumptions of the common factor model, PCA loadings tend to be too high whereas common factor loadings are very accurate.

4. Usefulness of assessing the percentage of explained common variance in exploratory factor analysis

The percentage of explained common variance can be used as a goodness of fit test, and this in itself is an important reason for computing it. However, some situations require this percentage to be computed: this is the case when Parallel Analysis (PA, Horn, 1965) is used to assess the dimensionality of a reduced correlation matrix.

PA was proposed by Dr. Horn to assess the dimensionality of a correlation matrix. The idea is to compare the eigenvalues of the empirical correlation matrix to eigenvalues of random correlation matrices: only the component associated with empirical eigenvalues larger than random eigenvalues should be retained. The size of the eigenvalues of a correlation matrix depends – among such other things as the size of the sample and the shape of the distribution of observed variables – on the number of variables in the correlation matrix: this means that, in order to be able to compare the empirical and the random eigenvalues, the random correlation matrices must be generated on the basis of the same number of observed variables. As the number of observed variables is the same, the empirical and the random correlation matrices have an identical amount of total variance: this is why eigenvalues from different correlation matrices can be compared. Although PA is usually based on the comparison of eigenvalues, this comparison could also be made in terms of the percentage of explained variance of each component (i.e., the eigenvalue divided by the total amount of variance).

When PA is used to assess the dimensionality of a reduced correlation matrix, the scenario is much more complex. The most important difficulty is related to the total amount of variance of the reduced correlation matrix (i.e., the amount of common variance). In a reduced correlation matrix the total amount of variance is not related to the number of observed variables: if random reduced correlation matrices are generated, each reduced correlation matrix will have a different total amount of common variance (which will probably be low because random variables are expected to be uncorrelated with one another), even if the same number of observed variables is always used. The consequence is that eigenvalues from different reduced correlation matrices cannot be compared with one another, because they are partitions of a different amount of common variance. The solution is to compare the percentage of explained common variance (instead of the eigenvalues). This is the key idea that we propose for assessing the dimensionality of a reduced correlation matrix with PA (see Timmerman, & Lorenzo-Seva, 2011).

References

Gorsuch, R.L.(1983). *Factor analysis* (2nd ed.) Hillsdale, NJ: Erlbaum.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179-185.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. Psychometrika, 23, 187-200.

Lorenzo-Seva, U. (2003). A factor simplicity index. *Psychometrika, 68*, 49-60.

Mulaik, S. A. (Ed.). (1992). Theme issue on principal components analysis. *Multivariate Behavioral Research, 27* (3).

Snook, S. C., & Gorsuch, R. L. (1989). Component analysis versus common factor analysis: A Monte Carlo study. *Psychological Bulletin*, *106*, 148–154.

Ten Berge, J. M. F., & Kiers, H. A. L. (1991). A numerical approach to the exact and the approximate minimum rank of a covariance matrix. *Psychometrika*, 56, 309-315.

Ten Berge, J. M. F., Snijders, T. A. B. & Zegers, F. E. (1981). Computational aspects of the greatest lower bound to reliability and constrained minimum trace factor analysis. *Psychometrika*, *46*, 201-213.

Thompson, B. (1992). A partial test distribution for cosines among factors across samples. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 2, pp. 81-97). Greenwich, CT:JAI Press.

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality Assessment of Ordered Polytomous Items with Parallel Analysis. *Psychological Methods*, *16*, 209-220.

Widaman, K. F. (1993). Common factor analysis versus principal component analysis: Differential bias in representing model parameters? *Multivariate Behavioral Research*, *28*, 263-311.

Woodhouse, B. & Jackson, P.H. (1977). Lower bounds to the reliability of the total score on a test composed of nonhomogeneous items: II. A search procedure to locate the greatest lower bound. *Psychometrika*, *42*, 579-591.