

「CROP」

**HOW MANY AXES IN YOUR FACTOR
ANALYSIS?**

THE LITTLE UNKNOWN HISTORY

DECEMBER 2010

WWW.CROP.CA



life to ideas

How many axes in your factor analysis? The little unknown history.

Thoughts on the number of axes to be kept in a factor analysis, or how obsolete economic constraints still affect factor analysis.

Abstract

Market research frequently uses factor analysis, especially in order to reduce the number of dimensions of the analysis space, get rid of residual noise, and set up the table for clustering or regression. A commonly used criterion to determine the number of axes to be kept is a minimum threshold (usually 1 or slightly above 1) to be applied to the associated eigenvalues – Kaiser Guttman rule (KG). This seems reasonable since a retained axis should carry at least as much information as every single variable on which the analysis space is built. What has always looked awkward to us, however, is that this criterion is applied before the axes rotation while the whole remaining analysis is performed after rotation. We will demonstrate through a few examples why it seems more logical to apply the criterion after rotation (in which case one does not talk about eigenvalues, but their equivalent: sum of squared loadings) and will tentatively suggest explanations of the current practice, one of which is rather unexpected. Finally, we shall shortly explain our own practice.

Introduction

Factor analysis (FA) is a method that is widely used in market research. In order to avoid confusion with the original FA in psychology, which aims at uncovering hidden factors behind the measured variables, the term *exploratory* FA is often used to describe a method whose objective is to summarize multidimensional data (battery of attitudinal items, satisfaction scores, etc.) in a small number of unambiguous factors, as distinct as possible one from another. One seeks simultaneously a synthesis (to extract sense), a simplification into more easily manageable variables (before a regression or a segmentation, for instance), and noise reduction, by discarding residual dimensions. In what follows, we shall refer to exploratory FA applied to market research. We assume that the reader has a basic practical knowledge of FA and its usual terminology. Since it is the common practice, let us assume a principal component analysis

(PCA), followed by a Varimax rotation, which seeks the best alignment between the axes and the original variables.

One of the first problems faced by the researcher performing a FA is the dimensionality of the factorial space, in other words the number of axes to be retained. This number is inferior or equal to the number of original variables, and should ideally be strictly inferior if we effectively wish to reduce the data. Although theoretically there is not yet an optimal solution, many rules are presented in the literature, of which we shall quote the 3 simplest and most commonly used:

1. The Kaiser-Guttman (KG) rule proposed in the 1950's-1960's, states that only the first axes (principal components) whose *eigenvalues* are superior or equal to 1 (or an arbitrary threshold slightly superior like 1.25) should be retained.
2. The analysis of the *scree plot*, a diagram showing the eigenvalues against each axis order number. One looks for an elbow, i.e. an important increase in the graph curvature.
3. An arbitrary predetermined threshold of the *proportion of variance* to be explained by the factorial space. One could, for instance, choose to keep the n first axes cumulating at least 75% of the total variance of the original variables.

The first criterion seems logical: any axis accounting for less information than a single variable is to be considered residual and therefore discarded. The second criterion assumes that when the graph of the scree plot becomes almost flat, it is an indication that any new axis does not contribute much additional information, relative to the previous ones.

However, what has never seemed logical to us is the fact that criteria 1 and 2 do not take into account the final analysis reference system, since they are applied before any axis rotation is applied. Nevertheless, all the remaining analyses are usually performed on rotated axes.

It is as if we answer the question of the number of axes of the principal component analysis (mathematical axes), and forget that we are indeed performing a factor analysis (factors or significant axes).

This problem does not concern criterion 3, because the explained variance will be the same in the space before and after rotation (same subspace with different coordinate systems).

In what follows, we shall try to demonstrate through simple examples what seems illogical about this practice, and dare to provide an unexpected historical explanation for it.

A few examples

Example 1 comes from a genuine FA on 20 attitudinal variables about a consumer product. According to the KG criterion, 6 axes should be kept. However, if we retain 7,8, and even 9 axes, we realize that, *after rotation*, the associated variance (sum of squared loadings, similar to the eigenvalue) is still greater than 1 for these 3 additional axes.

Their interpretation (which would be too lengthy to detail here) is also consistent and unambiguous. Why should we reject them? We have drawn the scree plot and its equivalent for the associated variances (see figure 1).

We first realize that the graph curvature after rotation is less important and that the graph lines after rotation meet the 1 mark further than the graph before rotation. It seems normal, since the variance of the first axes is redistributed on the following ones. Moreover, the graphs after rotation are less regular (zigzag). This is also normal because semantic via the rotation imposes sense to purely mathematical objects. It is this very semantic which distinguishes a FA from a PCA. Why ignore it?

Our second example (figure 2) demonstrates the very irregular pattern of the scree plot for rotated axes when compared to the regular scree plot. There is no chance that we find an elbow here if we were to apply criterion 2. This example also illustrates the fact that the more axes we have, the less the problem is relevant, since the graph of eigenvalues tends to meet the graph of variance of the rotated axes at the 1 mark. This is because the variance of the first axes is distributed on a larger number of factors and its contribution to the last axes becomes negligible.

Our last example is a fictitious extreme example intended to help understand the problem. One can easily reproduce it, with any statistical software package. In a survey database, let us create 2 random variables (say of uniform law), X and Y. Their correlation coefficient is close to 0. Let us define $Z = X + Y$.

Clearly, Z is correlated with X ($R \approx .7$). If we perform a FA on X and Z, using the regular KG criterion (eigenvalues ≥ 1) we obtain only one axis. On the other hand, if we force 2 axes, and impose a rotation on them, we

obtain 2 axes with variance 1. The dimensionality of our factorial space is therefore 2. A FA directly on variables X and Y gives similar results provided there exists a minute non 0 correlation between them. In this case, however, it can be objected that all prerequisites for a FA are not met (presence of correlation to start with).

This is an extreme example, on the borderline of philosophy. From the original FA standpoint, say in psychology, we could consider this factorial space as being of dimension 1, since we are looking for the subspace of common information. From the exploratory FA standpoint, and clearly from a geometrical standpoint, this space is of dimension 2.

Component	Total Variance Explained			Extraction Sums of Squared Loadings					
	Initial Eigenvalues			Unrotated		Rotation			
	Total	% of Variar	Cumulative	Total	10 axes	9 axes	8 axes	7 axes	6 axes
1	3.84	19.2	19.2	3.84	3.11	3.12	3.15	3.22	3.18
2	2.29	11.5	30.7	2.29	2.09	2.13	2.17	2.16	2.14
3	1.38	6.9	37.6	1.38	1.57	1.56	1.56	1.59	1.59
4	1.30	6.5	44.1	1.30	1.37	1.36	1.39	1.46	1.56
5	1.19	5.9	50.0	1.19	1.14	1.21	1.25	1.39	1.39
6	1.05	5.2	55.3	1.05	1.07	1.16	1.21	1.12	1.21
7	0.95	4.8	60.0	0.95	1.06	1.07	1.11	1.06	
8	0.88	4.4	64.4	0.88	1.04	1.06	1.04		
9	0.81	4.0	68.5	0.81	1.04	1.02			
10	0.76	3.8	72.3	0.76	0.97				
11	0.71	3.6	75.8						
12	0.70	3.5	79.3						
13	0.69	3.5	82.8						
14	0.64	3.2	86.0						
15	0.55	2.7	88.7						
16	0.53	2.7	91.4						
17	0.49	2.4	93.8						
18	0.48	2.4	96.2						
19	0.40	2.0	98.2						
20	0.35	1.8	100.0						

Extraction Method: Principal Component Analysis.

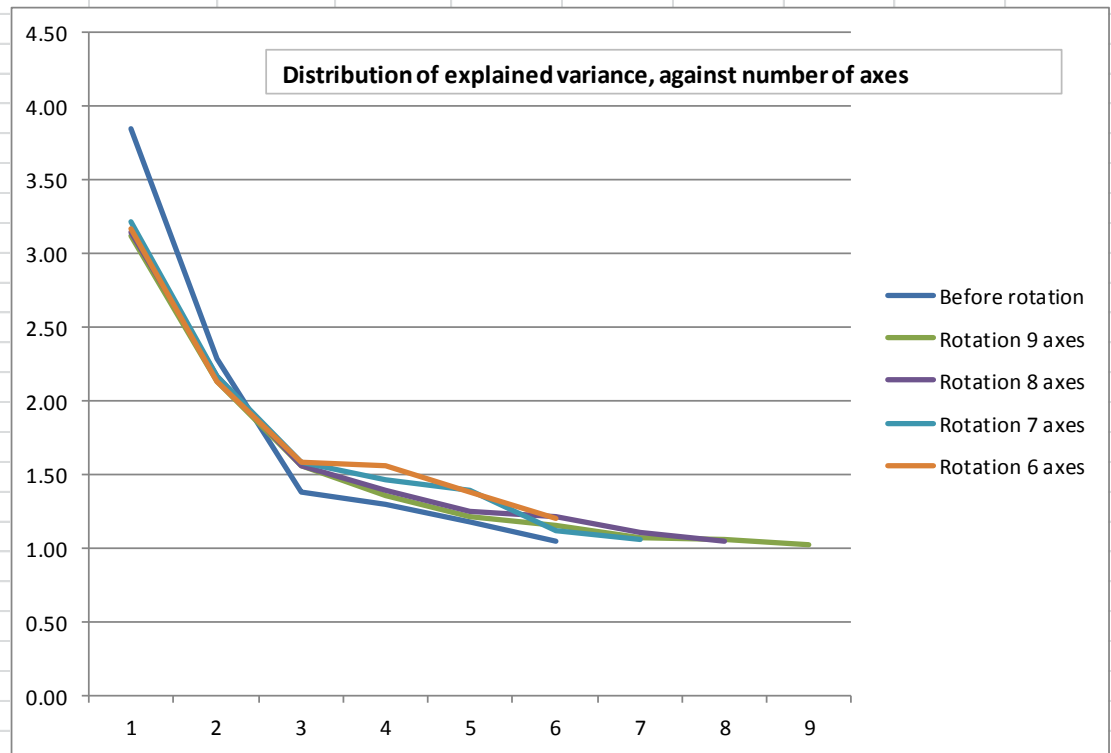


Figure 1

Total Variance Explained				Extraction Sums of Squared Load			Rotation Sums of Squared Load					
Component	Initial Eigenvalues			Total	% of Variance	Cumulative	Total	% of Variance	Cumulative	Total	% of Variance	Cumulative
	Total	% of Variance	Cumulative									
1	3.68	10.8	10.8	3.68	10.8	10.8	2.50	7.3	7.3			
2	2.57	7.6	18.4	2.57	7.6	18.4	2.20	6.5	13.8			
3	2.23	6.6	25.0	2.23	6.6	25.0	2.14	6.3	20.1			
4	1.96	5.8	30.7	1.96	5.8	30.7	1.99	5.9	26.0			
5	1.40	4.1	34.8	1.40	4.1	34.8	1.96	5.8	31.7			
6	1.30	3.8	38.7	1.30	3.8	38.7	1.64	4.8	36.6			
7	1.22	3.6	42.3	1.22	3.6	42.3	1.57	4.6	41.2			
8	1.15	3.4	45.7	1.15	3.4	45.7	1.28	3.8	44.9			
9	1.07	3.2	48.8	1.07	3.2	48.8	1.24	3.6	48.6			
10	1.03	3.0	51.8	1.03	3.0	51.8	1.08	3.2	51.8			
11	1.00	2.9	54.8	1.00	2.9	54.8	1.03	3.0	54.8			
12	0.98	2.9	57.7									
13	0.96	2.8	60.5									
14	0.94	2.8	63.3									
15	0.93	2.7	66.0									
16	0.91	2.7	68.7									
17	0.90	2.6	71.3									
18	0.87	2.6	73.9									
19	0.85	2.5	76.4									
20	0.83	2.4	78.8									
21	0.81	2.4	81.2									
22	0.72	2.1	83.3									
23	0.69	2.0	85.4									
24	0.68	2.0	87.4									
25	0.65	1.9	89.3									
26	0.58	1.7	91.0									
27	0.53	1.6	92.6									
28	0.51	1.5	94.1									
29	0.45	1.3	95.4									
30	0.43	1.3	96.6									
31	0.39	1.1	97.8									
32	0.31	0.9	98.7									
33	0.23	0.7	99.4									
34	0.21	0.6	100.0									

Extraction Method: Principal Component Analysis.

Figure 2

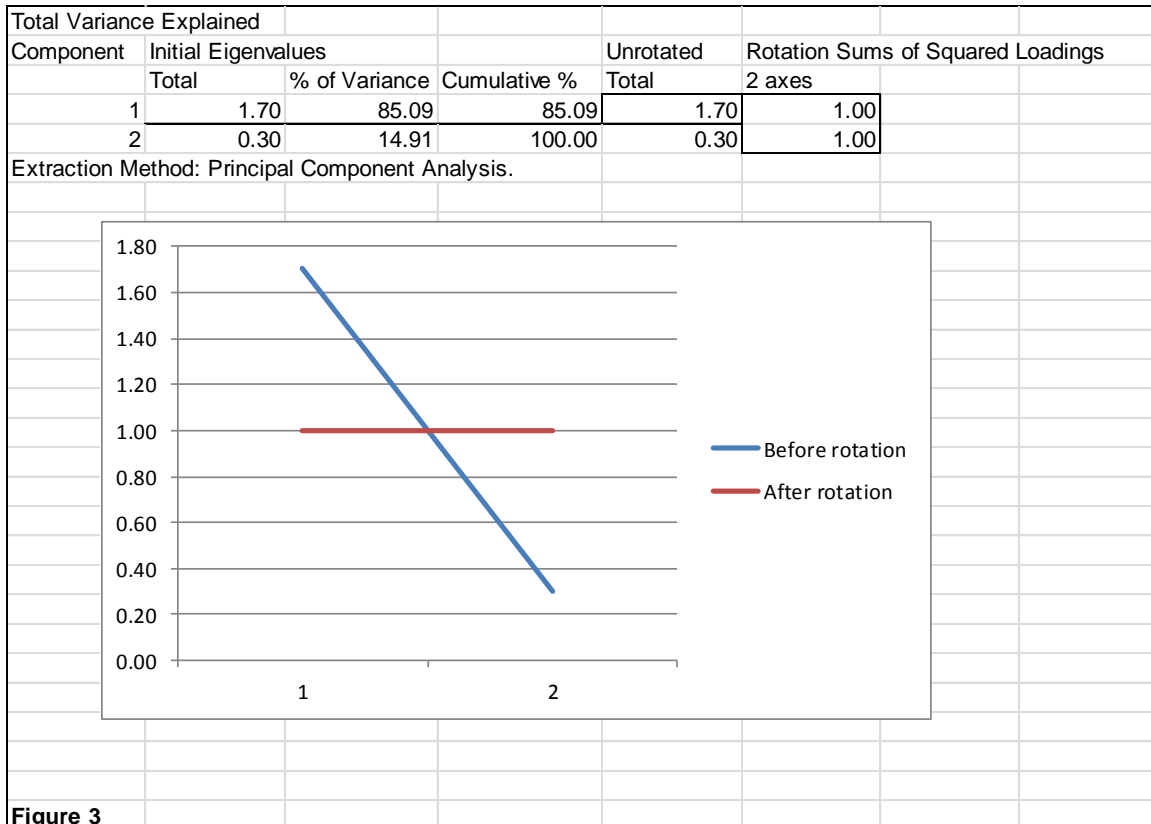


Figure 3

Some history

While writing this paper, we have found dozens of articles on the subject, offering sophisticated and elegant methods to determine the dimensionality of a factorial space, but we could not find one challenging this *dogma* that one first establishes the dimensionality of the space, and then operates a rotation. Before concluding, we wish to suggest 2 possible explanations of this fact.

First, recent research in this domain comes most often from artificial intelligence, data mining, and other disciplines more interested in performance than in sense. For instance, an interesting article from Mu Zhu and Ali Ghodsi (*Automatic dimensionality selection from the scree plot via the use of profile likelihood*, Waterloo 2005) suggests an efficient algorithm in the PCA context. This is an application in automated image recognition and if powerful algorithms are sought for large factorial spaces, one does not care much about interpreting the resulting axes. The essence here is noise reduction.

But the main reason seems historical: two pionniers of FA were Kaiser and Guttman, in the 1950's and 1960's. In those days, computer time was

very expensive – when one did not compute by hand! Axes rotation is computation intensive where the number of operations is of magnitude $n!$ if n is the number of original variables. A single rotation could cost hundreds of dollars. In this context, applying a threshold to the eigenvalues looks like a clever economical trick. Otherwise, in order to apply the equivalent of the KG criterion after rotation, one has to perform a rotation of the coordinate system for all integers greater than the number of axes found by KG until the variance of the last resulting axis becomes less than 1. (in example 1, one has to do a rotation on 7, 8, 9, and 10 axes before establishing that the threshold is reached at 9). What can be done in a couple of seconds or milliseconds today represented a huge task 50 years ago. We are thus living with the obsolete legacy of a not so remote era. Why the industry did not adapt is hard to understand. Inertia, lack of historical hindsight? For us, it is the perfect example of the syndrome: “My stat package does it this way, it should be right”. A question for the skeptical: Why is it that in 2010 the default maximum number of iterations for a rotation in the most used and taught statistical package is still 25, just like in 1970. This number is often too small, and it could be increased to 10,000 without anyone complaining. Imagine the number of students and researchers who have to adjust it everyday!

A new rule: KG+

Our practice in FA is close to KG (the threshold depends on the context – the more variables there are, the higher the threshold). If KG establishes n axes, we usually look also at solutions with $n+1$, $n+2$, $n+3$ etc. axes, comparing the additional axis variance *after rotation* with the threshold. In most cases, these axes are interpretable and make sense.

From our perspective, KG is used to establish a *minimum* number of axes. KG applied to the rotated axes, let us call it KG+, helps us determine the number of significant axes or factors, which often slightly exceeds the number obtained with KG. The ultimate criterion remains that the axes must make sense. We believe that the problem has not been fully solved, and that current technology should help new innovative solutions to emerge.

To come

In an article to come we shall explain why we think that is also important to look at the unrotated solution, even though most of its axes do not make sense.

References

The 2 following articles contain a relevant and extensive bibliography on the subject of establishing the number of factors in a FA : *Automatic dimensionality selection from the scree plot via the use of profile likelihood* (Waterloo 2005) by Mu Zhu and Ali Ghodsi quoted above and *Finding the Magic Number* (The Psychologist, October 2008) by Paul Wilson and Colin Cooper.

Acknowledgment

We wish to thank Michel Saulnier of the MRIA for his comments and corrections on the first version of this article.

「CROP」

「

life to ideas

550, RUE SHERBROOKE OUEST
MONTRÉAL (QUÉBEC) H3A 1B9

BUREAU 900 – TOUR EST

T 514 849-8086, POSTE 3064