EXPLORATORY FACTOR ANALYSIS IN MPLUS, R AND SPSS

Sigbert Klinke[1,2] Andrija Mihoci[1,3] and Wolfgang Härdle[1,3]
[1]School of Business and Economics , Humboldt-Universität zu Berlin, Germany
[2]Department of Law and Economics, Johannes-Gutenberg-Universität Mainz, Germany
[3]Humboldt-Universität zu Berlin, C.A.S.E. Center for Applied Statistics and Economics, Germany
sigbert@wiwi.hu-berlin.de

In teaching, factor analysis and principal component analysis are often used together, although they are quite different methods. We first summarise the similarities and differences between both approaches. From submitted theses it appears that student have difficulties seeing the differences. Although books and online resources mention some of the differences they are incomplete. A view either oriented on the similarities or the differences is reflected in software implementations. We therefore look at the implementations of factor analysis in Mplus, R and SPSS and finish with some conclusions for the teaching of Multivariate Statistics.

INTRODUCTION

Exploratory factor analysis (EFA) and Principal component analysis (PCA) are integral and important parts in every lecture on multivariate statistics. We teach PCA at Humboldt-Universität zu Berlin usually first in Multivariate Statistical Analysis I; and EFA as well in Multivariate Statistical Analysis I and II, once for metrical data and once for ordinal data. Although EFA of ordinal data is based on the EFA of metrical data (or at least the underlying variable approach is) and therefore is a repetition of metrical EFA, most of our students have shown severe difficulties in distinguishing between both techniques in their theses.

We believe that three factors are mainly responsible for this:

- the attitude, knowledge and experience of the teacher,
- the way the techniques and their differences are treated in statistics books and online resources, and
- the way that EFA and PCA are implemented in software.

PRINCIPAL COMPONENT ANALYSIS AND EXPLORATORY FACTOR ANALYSIS

Principal component analysis

The idea of PCA is the representation of a high-dimensional dataset by a linear low-dimensional subspace. This achieved by an orthogonal rotation of the coordinate system (or an orthogonal rotation of the dataset depends on your personal view). The optimisation criterion, the total variance which is the sum of the squared L_2-distances between the data points and the centre of the data, is invariant under rotation. Therefore finding a low-dimensional subspace which maximises the percentages of incorporated total variance is equivalent to preserving as much as possible of the original distances between the data centre and the data points in the low-dimensional linear subspace. Since rotations are used, the link between the principal components and the original variables PCA turns out to be a linear model and the low-dimensional subspace can be described by the eigenvectors of the covariance or correlation matrix with the largest eigenvalues. The eigenvalues and vectors are computed by Singular Value Decomposition of the covariance or correlation matrix.

Exploratory factor analysis

The idea of EFA is to model a set of variables by latent factors. The model used is a linear model, but in order to perform the EFA we need additional model assumptions, e.g.,

1. The factors have zero mean and unit variance and are uncorrelated,
2. The error terms have zero mean and are uncorrelated, and
3. The factors and error terms are uncorrelated.

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

But even these assumptions allow for a variety of different methods in the practical use of EFA which make this method very useful in applied work (compare Pett, Lackey & Sullivan 2003, p. 7). Usually four steps are done to perform EFA:

1. Estimation of the correlation or covariance matrix.
2. Estimation of the number of common factors (Elbow criterion, Kaiser criterion, Parallel analysis, etc.).
3. Estimating the loadings matrix of the common factors (Principal component, principal axis, maximum-likelihood, unweighted least squares ,etc.) and
4. Rotation of the loadings to improve the interpretability of the factors (Varimax, Quartimax, Promax, etc.).

Similarities and differences

A first look reveals a lot of similarities between both techniques:

- Reduction of the number of variables,
- Can be used on covariance or correlation matrix,
- Based on a linear model,
- Similar results for the resulting principal components and latent factors (without rotation).

However a deeper look reveals (subtle) differences:

- PCA starts with the search for the best fitting (in a L_2 sense) linear low-dimensional subspace and ends up with a linear model; therefore a descriptive technique. EFA directly builds up on a linear model; therefore a model-based technique.
- PCA has an importance ranking of the components determined by the eigenvalues whereas in EFA we first choose the dimension of the factor space and later an appropriate coordinate system such that the factors are all equal in the analysis.
- PCA violates the assumption of uncorrelated error terms since the error terms are linear combinations of the eigenvectors which do not belong to the selected linear low-dimensional subspace.
- PCA optimises the total variance. Since the total variance is the sum of squared distances to the data centre it is obvious that the covariance or correlation structure of the data does not play any role. EFA aims to reproduce the covariance or correlation matrix as *well* as possible.

STUDENTS

Twelve students from different Master programmes (half of them from the Joint Master Programme in Statistics) handed in project theses for the lecture Multivariate Statistical Analysis II. The aim of their project thesis was to apply the multivariate methods they learned in Multivariate Statistical Analysis II (and I) lectures to a self-selected dataset. Since the lectures did not involve any practical work, the idea was that they apply the methods practically.

One student did not use PCA or EFA at all, two students did not apply PCA, one student did not apply EFA and the rest applied both techniques. Half of the students used R (R Development Core Team, 2009), three students used SPSS, one student used both and one student used Stata. None of the students used PCA or EFA based on the covariance matrix of the variables. Most of the students used only one method for EFA (principal component, principal axis or maximum-likelihood) and only one student compared and discussed four extraction methods. All of the students using R used the `factanal` function which is based on the maximum-likelihood method. The students either did not check the assumption, the multivariate normality of the data, or applied it although the data were non-normal. One student even argued with the results of the built-in test of `factanal`. Only the student who discussed the four extraction methods excluded the maximum-likelihood method because of the formerly shown non-normality of the used data. All of

the students who used SPSS analysed the Measure of adequacy and two applied the Bartlett test although the non-normality of their data was not tested.

Two students who used PCA and EFA tried to discuss the difference between both techniques but covered only the first point (descriptive vs. model-based) mentioned in the differences subsection above. One student, under the heading, *Principal component analysis* produced a Varimax rotation of the components.

The students who applied both techniques ran into difficulties when starting to interpret the results from PCA and EFA, especially when they used SPSS. For example, some students avoided giving an interpretation for PCA. Obviously these students did not care much about the differences between PCA and EFA. This raises the question: *Although they ran into problems why do most of them not seem to see the differences between both techniques?*

TEACHERS AND LITERATURE

Teachers with a mathematical perspective, in particular, do not value the EFA *because of* the subjective choices. Teachers with a more applied perspective value the method *because of* the subjective choices. The discomfort about EFA stems from the fact that two people with the same data could come to different results. Although in our practical work, e.g. in educational science, the different choices of the extraction or rotation method, in contrast to the methodological differences, does not matter that much. On the level of interpreting the factors, the results are the same. The situation is different when we are talking about the number of factors; here only theoretical knowledge about the subject involved, outside of statistics, can help.

In PCA we have a clear defined algorithm and any two people applying it to the same data will compute the same eigenvalues and eigenvectors. Only the number of important principal components may differ. However, interpreting the principal components is often very difficult.

A comparison of the books of Härdle and Simar (2003) *Applied multivariate statistical analysis;* Bartholomew, Steele, Moustaki and Galbraith (2002) *The analysis and interpretation of multivariate data for social scientists* and the entries in the German Wikipedia show a clear difference:

- The book by Härdle and Simar uses 40 pages to explain PCA and 25 pages to explain EFA. The book of Bartholomew et al. uses 37 pages to explain PCA and 32 pages for the EFA. The entries *Hauptkomponentenanalyse* (PCA) and *Faktorenanalyse* (EFA) in the German Wikipedia have six and four-and-a-half pages, respectively.
- The difference between PCA and EFA is mentioned in the book by Härdle and Simar on half a page somewhere within the EFA chapter. The book by Bartholomew et al. uses ¾ of a page to emphasise the differences at the beginning of the EFA chapter, an accentuated place within the chapter. In the German Wikipedia it is even worse, only a ¼ of a page, a section in the middle of the *Faktorenanalyse* entry, is used to explain the differences between PCA and EFA.

PCA carries in both books and the German Wikipedia more weight which can be seen by the number of pages used to explain it. Also, neither of the two books and the Wikipedia entry mentions all four of the differences given above. Since the books reflect the attitude of the authors, who are all teachers, then how can students get a clear picture of the difference between PCA und EFA?

SOFTWARE

In our courses on Multivariate Statistical Analysis I and II we use different software to illustrate the concepts: among others packages are SPSS, R, and Mplus. Which views do the software packages take on PCA and EFA?

SPSS

It does not offer the PCA program as a separate menu item somewhere in the menu (and neither in the SPSS language). The PCA program is integrated into the EFA program which can be found under *Analyze → Data reduction → Factor*. It is obvious that in SPSS the similarities

between PCA and EFA are clearly emphasised. SPSS can base EFA on the covariance and correlation matrix, but only for metric variables. Since with the SPSS language correlation matrix can also be given to the `FACTOR` command an analysis of a polychoric correlation matrix is possible. It offer a range of methods in EFA to select the number of factors, extraction and rotation methods (see Table 1).

Mplus

The User's Guide title is *Mplus – Statistical analysis with latent variables;* therefore no support for PCA is given. The EFA is based on a correlation matrix which depends on the type of variables (metric, categorical and a mixture of both). The number of factors (from-to) needs to be given explicitly in the modelling step and Mplus computes all necessary factor models in parallel. It provides for each model performance numbers such as the Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA) or Standardised Root Mean Square Residual (SRMR). It offers a great variety on extractions (*neither* principal component *nor* principal axis) and rotation methods. As a consequence Mplus takes a firm view on pure EFA avoiding any PCA or PCA related method.

R

Since R is a programming language with a large number of additional packages it offers several packages for EFA:

- `stats` (R Development Core Team, 2009) is a standard package installed with R. It offers separate functions for PCA (`princomp`) and EFA (`factanal`). In EFA only the maximum-likelihood method is supported. Two rotation methods as separate functions are available: `varimax` and `promax`. Again we find a clear separation of PCA and EFA.
- `efa` (Klinke & Wagner 2008) is a collection of function to perform for EFA like in SPSS. However, it offers less extraction and rotation methods, but aims to compute several factor models in parallel like Mplus.
- `FactoMineR` (Husson & Josse & Le & Mazet 2009) is an R package for exploratory data analysis. It offers both PCA and EFA and they are handled within the same framework, the *multiple factor analysis* (Escofier & Pagès 1994). For continuous variables it uses PCA to perform EFA and for categorical variables multiple correspondence analysis. Rotations are provided via the `stats` package.

There are three more specialised packages, which extend the EFA or mix it with other techniques:

- `FAiR` (Goodrich 2009) estimates factor analysis models using a genetic algorithm, which permits a general mechanism for restricted optimisation with arbitrary restrictions.
- `FactoClass` (Pardo & del Campo 2009) combines factor and cluster analysis.
- `ifa` (Viroli 2007) is package for independent factor analysis.

Most packages, except our own collection of `efa` functions, extend the functionality of EFA which either does not allow a direct inclusion of PCA or uses approaches which are not directly comparable with the standard EFA.

We notice that software such as SPSS and Mplus take different views on PCA and EFA, SPSS emphasising the similarities and Mplus more the dissimilarities. In R we can find both views: in the basic package `stats` a clear separation and in the `efa` functions an SPSS like view.

Table 1. Overview of the features of different software for EFA

Software	Mplus 5.2.1	R 2.10.0[1]			SPSS 17
Package		stats	FactoMineR	efa	
Measurement of relationship					
Covariance	N	Y	Y	0[2]	Y
Correlation	Y	Y	Y	Y	Y
Polychoric	Y	N	Mixed	Y	0[3]
Selection criteria					
Scree Plot / Elbow	N	N		Y	Y
Horn	N	N		Y	0[4]
% Explained Var.	N	Y		Y	Y
Further criteria	TLI, CFI RMSEA, SRMR	N		N	N
Method					
PCA	N	Y	Y	Y	Y
ML	Y[5]	Y	N	Y	Y
Principal axis	N	N	N	Y	Y
ULS	Y[5]	N	N	Y	Y
GLS	Y	N	N	N	Y
Further methods	Weighted Least Squares[5]	N	N	N	Alpha and image factoring
Rotation					
Varimax	Y	Y	N	Y	Y
Promax	Y	Y	N	Y	Y
Further methods	Quartimax Quartimin Oblimin Crawfer[6] Geomin Equamax	N	N	N	Quartimax Oblimin Equamax

[1]For the packages in the software R an *N* means that the package does not directly incorporate such a function or parameter, but of course it can be part of another package.
[2]This feature has not been thoroughly tested.
[3]The SPSS command FACTOR allows for a correlation matrix as a parameter to be given.
[4]An SPSS program can be downloaded from https://people.ok.ubc.ca/brioconn/nfactors/nfactors.html.
[5]Mplus offers a wide range of extraction methods with minor differences (Muthén & Muthén 2007, p. 482ff).
[6]Mplus offers a range of rotation methods based on the so-called Crawford-Ferguson family (Browne 2001).

CONCLUSION

A large part of our master students, even from Master of Statistics, do not have a clear view of the differences between PCA and EFA as their project theses prove. All of the sources available to them: teachers, their books, online resources and the software used, all handle these aspects differently. In the books the differences are given, but not always prominently. The software takes both views; a clear separation of both techniques like in Mplus or in the R package `stats`, and a mix of it as in SPSS and the R functions of `efa`.

As teachers we have to clarify the differences theoretically as well as with practical examples and it would be a good idea to use appropriate software such as Mplus or the R package `stats` to make the difference between both techniques visible.

REFERENCES

Bartholomew, D. J., & Steele, F., Moustaki, I., & Galbraith, J. I. (2002). *The analysis and interpretation of multivariate data for social scientists.* Chapman & Hall.

Browne, M. W. (2001). An overview of analytic rotations in exploratory factor analysis. *Multivariate Behavioral Science, 36,* 111-150.

Escofier B., & Pagès J. (1994). Multiple factor analysis (AFMULT package). *Computational statistics & data analysis, 18,* 121-140.

Härdle, W., & Simar, L. (2003). *Applied multivariate statistical analysis.* Berlin: Springer Verlag.

Goodrich, B. (2009). FAiR: Factor analysis in R (version 0.4-4). The Comprehensive R Archive Network, http://cran.r-project.org/web/packages/FAiR/ (accessed 30 Oct 2009).

Husson, F., Josse, J., Le, S., & Mazet, J. (2009). FactoMineR: Factor analysis and data mining with R. Rennes/France: The Comprehensive R Archive Network, http://factominer.free.fr and http://cran.r-project.org/web/packages/FactoMineR (both accessed 30 Oct 2009).

Klinke, S., & Wagner, C. (2008). Visualizing exploratory factor models. *COMPSTAT 2008 - Proceedings in Computational Statistics - 18th Symposium held in Porto (Portugal)* by P. Brito (ed.). Heidelberg: Physika Verlag (CDROM), Software: http://stirner.wiwi.hu-berlin.de/mediawiki/mmstat de/index.php/Spezielle Themen - R-Visualisierung von Faktormodellen (accessed 30 Oct 2009).

Muthén, L. K., & Muthén, B. O. (2007). *Mplus: Statistical analysis with latent variables - User's guide (fifth edition).* Los Angeles, CA: http://www.statmodel.com (accessed 30 Oct 2009).

Pardo, C. E., & del Campo, P. C. (2009). FactoClass: Combination of factorial methods and cluster analysis (version 1.0.1). The Comprehensive R Archive Network. Retrieved 30 October 2009 from: http://cran.r-project.org/web/packages/FactoClass.

Pett, M. A., Lackey. N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis.* Thousand Oaks, CA: Sage.

R Development Core Team (2009). *R: A language and environment for statistical computing*, Vienna, Austria: R Foundation for Statistical Computing.

Viroli, C. (2007). ifa: Independent factor analysis (version 5.0). Bologna/Italia: The Comprehensive R Archive Network. Retrieved 30 October 2009 from: http://cran.r-project.org/web/packages/ifa/.