Finding Structures of Interest in a Large Data Set Using Factor Analysis

Peter Filzmoser
Department of Statistics and Probability Theory
Vienna University of Technology, AUSTRIA

Abstract: In this paper we introduce a statistical method which can be used in combination with principal component analysis or factor analysis. Certain variables of a large data set which are of interest can be selected in order to calculate loadings and scores of these variables. We describe how the remaining variables of the data set can be presented in the previously extracted factor space. Furthermore, a possibility for the representation of the results is shown which is helpful for the interpretation.

Zusammenfassung: In diesem Artikel stellen wir eine statistische Methode vor, die in Kombination mit Hauptkomponenten– oder Faktorenanalyse angewandt werden kann. Bestimmte Variablen eines großen Datensatzes, die von Interesse sind, können ausgewählt werden, um Ladungen und Faktorenwerte dieser Variablen zu berechnen. Wir beschreiben, wie die restlichen Variablen dieses Datensatzes im vorher extrahierten Faktorraum dargestellt werden können. Weiters wird eine Möglichkeit zur Präsentation von Ergebnissen gezeigt, die für die Interpretation hilfreich ist.

Keywords: Principal component analysis, Factor analysis, Biplot, Projection.

1 Introduction

Sometimes we are interested in special variables of a possibly large data set. Suppose we want to know the relations of these special variables between each other but also to the other variables of the data set. If, in this case, we performed principal component analysis (*PCA*) or factor analysis (*FA*) onto the whole data set, the solution would possibly not allow a good interpretation for the variables of special interest, because the variation of these variables is often too low in comparison to the total variation.

A solution for this task is to perform *PCA* or *FA* only onto the variables of special interest. Loadings and scores can be calculated at the basis of the correlation matrix of these special variables. The results are new axes or factors which will describe the selected variables rather good, since only the variation of these variables has been considered.

In order to find out the relations to the other variables of the data set, these remaining variables can be projected onto the factors extracted before. This means that the loadings of the remaining variables in the factor space arising from the selected variables have to be determined (Section 2).

In Section 3 we consider the (usual) case that the results are presented in lower-dimensional projections of the factor space (e.g. projections onto pairs of factors). If relations between the variables are to be investigated, the interpretation of the results can be facilitated, if only variables are projected which are close to the lower-dimensional factor space, where a measure of closeness has to be defined.

In Section 4 we show an example with a large data set. We emphasize both the calculation with the introduced method and the interpretation of the results.

2 Calculation of the Loadings

Let us consider a data matrix $\boldsymbol{X}$ $(n \times p)$ with p variables and sample size n. Without loss of generality we assume that the first s variables are of special interest. Therefore we select the first s variables and denote this sub-matrix by $\boldsymbol{X}_s$. The remaining $p - s = r$ variables are combined in the matrix $\boldsymbol{X}_r$. With the notation $\boldsymbol{X} = (\boldsymbol{X}_s \boldsymbol{X}_r)$ in the following we denote the partition of $\boldsymbol{X}$ into two sub-matrices.

Standardizing the variables to mean zero and unit variance defines the matrices $\boldsymbol{Y}_s$ and $\boldsymbol{Y}_r$. Since each variable is standardized separately, we have $\boldsymbol{Y} = (\boldsymbol{Y}_s \boldsymbol{Y}_r)$, where $\boldsymbol{Y}$ is the matrix resulting from standardizing the whole data matrix $\boldsymbol{X}$.

The empirical correlation matrix of the selected variables is calculated by $\boldsymbol{R}_s = \frac{1}{n-1}\boldsymbol{Y}_s^\top \boldsymbol{Y}_s$, the correlation matrix of the remaining variables is $\boldsymbol{R}_r = \frac{1}{n-1}\boldsymbol{Y}_r^\top \boldsymbol{Y}_r$. The empirical correlation matrix of the whole data set is

$$\boldsymbol{R} = \frac{1}{n-1}\boldsymbol{Y}^\top \boldsymbol{Y} = \begin{pmatrix} \boldsymbol{R}_s & \boldsymbol{R}_{sr}^\top \\ \boldsymbol{R}_{sr} & \boldsymbol{R}_r \end{pmatrix} . \tag{1}$$

Basis of *PCA* or *FA* is the calculation of eigenvalues and eigenvectors of the correlation matrix (or more generally of a dispersion matrix). Clearly, the eigenvalues and eigenvectors of a correlation matrix in general are different from the eigenvalues and eigenvectors of a sub-matrix of this correlation matrix ($\boldsymbol{R}_s$ is a sub-matrix of $\boldsymbol{R}$), compare Mardia et al. (1979).

Let $\widehat{\boldsymbol{\Lambda}}$ denote the estimated loadings when using the correlation matrix $\boldsymbol{R}$ of all variables, and $\widehat{\boldsymbol{\Lambda}}_s$ the estimated loadings for $\boldsymbol{R}_s$. Since the loadings are calculated from the eigenvalues and eigenvectors of the dispersion matrix, we note with the above considerations that $\widehat{\boldsymbol{\Lambda}}_s$ is no sub-matrix of $\widehat{\boldsymbol{\Lambda}}$. In the usual case we also have the following: Since $\boldsymbol{R}_s$ is a $(s \times s)$-matrix with $s < p$, only $k_s \leq k$ factors have to be extracted, where k is the number of extracted factors of the whole correlation matrix $\boldsymbol{R}$. This means that $\widehat{\boldsymbol{\Lambda}}_s$ is a $(s \times k_s)$-matrix where $\widehat{\boldsymbol{\Lambda}}$ is a $(p \times k)$-matrix.

The factor analytical model for s selected variables is in the sample case

$$\boldsymbol{Y}_s = \boldsymbol{F}_s \boldsymbol{\Lambda}_s^\top + \boldsymbol{E}_s , \tag{2}$$

compare Harman (1967). $\boldsymbol{F}_s$ is a $(n \times k_s)$-matrix of factor scores, and $\boldsymbol{E}_s$ is the error matrix. For the estimation of the loadings $\boldsymbol{\Lambda}_s$ there are a lot of different methods like principal factor analysis, maximum likelihood method, canonical factor analysis, ... (see e.g. Basilevsky, 1994). The factor scores can be estimated for example by multiple regression analysis in the way

$$\widehat{\boldsymbol{F}}_s = \boldsymbol{Y}_s \boldsymbol{R}_s^{-1} \boldsymbol{\Lambda}_s \boldsymbol{\Phi}_s \tag{3}$$

(see e.g. Lawley and Maxwell, 1971). $\boldsymbol{\Phi}_s$ is the empirical correlation matrix of the factor scores.

In the introduction of this paper we mentioned that we want to project the remaining variables onto the space spanned by the extracted factors. This means that we have to consider the (sample) model

$$\boldsymbol{Y}_r = \boldsymbol{F}_s \boldsymbol{\Lambda}_r^\top + \boldsymbol{E}_r \ . \tag{4}$$

$\boldsymbol{Y}_r$ and $\boldsymbol{F}_s$ are known. The loadings for the remaining variables, $\boldsymbol{\Lambda}_r$, have to be estimated. This model may be seen as a regression model where $\boldsymbol{\Lambda}_r$ are the unknown regression coefficients. Minimization of the sum of the squared residuals gives

$$\widehat{\boldsymbol{\Lambda}}_r^\top = (\boldsymbol{F}_s^\top \boldsymbol{F}_s)^{-1} \boldsymbol{F}_s^\top \boldsymbol{Y}_r = \frac{1}{n-1} \boldsymbol{\Phi}_s^{-1} \boldsymbol{F}_s^\top \boldsymbol{Y}_r \ , \tag{5}$$

and thus the estimation for the unknown parameters is

$$\widehat{\boldsymbol{\Lambda}}_r = \frac{1}{n-1} \boldsymbol{Y}_r^\top \boldsymbol{F}_s \boldsymbol{\Phi}_s^{-1} \ . \tag{6}$$

By this estimation it is possible to project all variables into the space of the factors resulting from the selected variables.

If the estimation for the factor scores (3) is inserted into (6), we get

$$\widehat{\boldsymbol{\Lambda}}_r = \frac{1}{n-1} \boldsymbol{Y}_r^\top \boldsymbol{Y}_s \boldsymbol{R}_s^{-1} \boldsymbol{\Lambda}_s \boldsymbol{\Phi}_s \boldsymbol{\Phi}_s^{-1} = \boldsymbol{Y}_r^\top \boldsymbol{Y}_s (\boldsymbol{Y}_s^\top \boldsymbol{Y}_s)^{-1} \boldsymbol{\Lambda}_s \ . \tag{7}$$

This means, once we have calculated the loadings of the selected variables $\boldsymbol{\Lambda}_s$, we just need the (standardized) matrices of the selected and the remaining variables to calculate the loadings of the remaining variables in the common factor space.

If the number of factors k_s is larger than two, a representation of the result may be obtained by projecting variables and/or scores onto the plane spanned by different pairs of factors. The loadings of the variables in the plane are the corresponding columns of the matrix of loadings calculated with the help of (7).

3 Projection

It is quite common to represent the results in planes spanned by all different pairs of factors. The reason is that two-dimensional presentations are easy to survey although by this projection to two dimensions we loose information. Especially in the case of data sets where relations between the variables are to be investigated, care has to be taken with the interpretation of results arising from lower-dimensional projections.

If the number of extracted factors, k_s, is larger than two and the results are shown in planes, only the loadings of the variables at the corresponding factors spanning the plane are considered. The distances and angles between the projected variables in general are distorted, and this may result in misinterpretations.

In view of the interpretation of the result it makes sense to project only the variables which are "close" to the plane spanned by the chosen factors. Closeness means in this context that the loadings of the variables on the remaining factors are "small". In this case the distortion of the projection is low.

Now the weak formulations above have to be clarified. As a measure of closeness we consider the squared multiple correlation (*SMC*) coefficient between each variable and two factors.

Let y_i ($i \in \{1, \ldots, p\}$) be a (standardized) variable and $\boldsymbol{f} = (f_a, f_b)^\top$ two different factors ($a, b \in \{1, \ldots, k_s\}$; $a \neq b$). Then the *SMC* between y_i and $\boldsymbol{f}$ is defined as

$$\rho^2_{y_i,\boldsymbol{f}} = \boldsymbol{\rho}^\top_{\boldsymbol{f}y_i} \boldsymbol{\rho}^{-1}_{\boldsymbol{ff}} \boldsymbol{\rho}_{\boldsymbol{f}y_i} \tag{8}$$

(see e.g. Mardia et al., 1979). $\boldsymbol{\rho}_{\boldsymbol{f}_{y_i}}$ is the correlation matrix between $\boldsymbol{f}$ and y_i, and $\boldsymbol{\rho}_{\boldsymbol{ff}}$ is the correlation matrix of the factors f_a and f_b, which is the identity for orthogonal factors.

With this measure of closeness between variables and plane we can specify a certain bound (e.g. 0.5). Variables with a lower *SMC* coefficient than this bound are not projected, the other variables are presented in the result.

Especially for biplot representations (Gabriel, 1971) where relations between the variables and relations between variables and objects are shown, this procedure is advisable.

4 Example

We consider a data set with more than 800 variables from the fields economy, ecology, society, health, environment, and others. The variables were measured in the 96 Bavarian districts and cities, mainly in the year 1987.[1] For detailed information concerning this data set we cite a technical report of STUDIA and ALBTUM (1993).

This large and extensive data set stimulates a lot of questions, which are to be answered by statistical methods. One field of interest is to find out the reasons for the appearance of stomach cancer. Let us suppose that this disease is connected with the rate of unemployment. We select all variables which are relatively high correlated ($|r| \geq 0.6$) with the variables (cause of death) stomach cancer (of female persons) and unemployment. The correlations can be seen from Table 1. A description of these variables will be presented in Section 4.1.

For the further investigation we select all variables given in Table 1. With these variables we perform principal factor analysis selecting two factors. The resulting loadings

[1] We like to thank the Austrian research institute STUDIA in *Schlierbach* for placing these data at our disposal.

Table 1: *Stomach cancer (female)* and *rate of unemployment* with relatively highly correlated variables ($|r| \geq 0.6$)

stomach cancer (female)	r	unemployment	r
payment/salary	0.62	payment/salary	0.61
married/divorced	0.60	pop % tn.college	-0.66
assignmt2municip	0.61	farm 3-5 cows%	0.62
tax/inhabitant	-0.60	assignmt2municip	0.63
stomach cancer m	0.77	rent/sq.meter	-0.65

are rotated with the quartimax method. So we obtain the rotated loadings and factor scores of the selected variables. With the help of (7) we are able to calculate the loadings of the remaining variables in the two-dimensional factor space which are the co-ordinates for projection.

For interpretational purposes we only project variables with a *SMC* coefficient of more than 0.5 with the plane spanned by the two factors. A biplot representation of this plane with the most important variables is shown in Figure 1. The cosine between two variable vectors approximates the correlation, the projection of the objects onto the variable vectors approximate the data values (Gabriel, 1971).

4.1 Interpretation of Figure 1

The 96 Bavarian districts and cities are abbreviated with three letters (for cities, the third sign is a point). The abbreviation of the variables is explained in the following.

In the east of the diagram we have regions, mainly districts close to cities, which are characterized by success: a large portion in people with a technical college (`pop % tn.college`), the difference of immigration and migration (`immigrat-migrat`) (measured in a period of five years) is positive, high rents (`rent/sq.meter`) (per square meter), a large portion in people with secondary school (`second.school%`), university (`university%`), and grammar school (`grammar school%`), high estate prices (`estate price`), a large portion of employees (`empl % employee`) and employment in credit institutes and insurances (`empl % cred+insu`), high tax per inhabitant (`tax/inhabitant`).

In the opposite directions we have the poor regions with low education: a large portion of people in apprenticeship (`empl % apprentic`), a large portion in persons with primary school (`primary school%`), stomach cancer of female (`stomach cancer f`) and male people (`stomach cancer m`), a large portion in workers (`empl % worker`), a large portion in small farms (with 3 to 5 cows) (`farm 3-5 cows%`), high assignments to the municipalities (`assignmt2municip`), a high relation of payment per salary (`payment/salary`), and a high rate in unemployed people (`unemployment`).

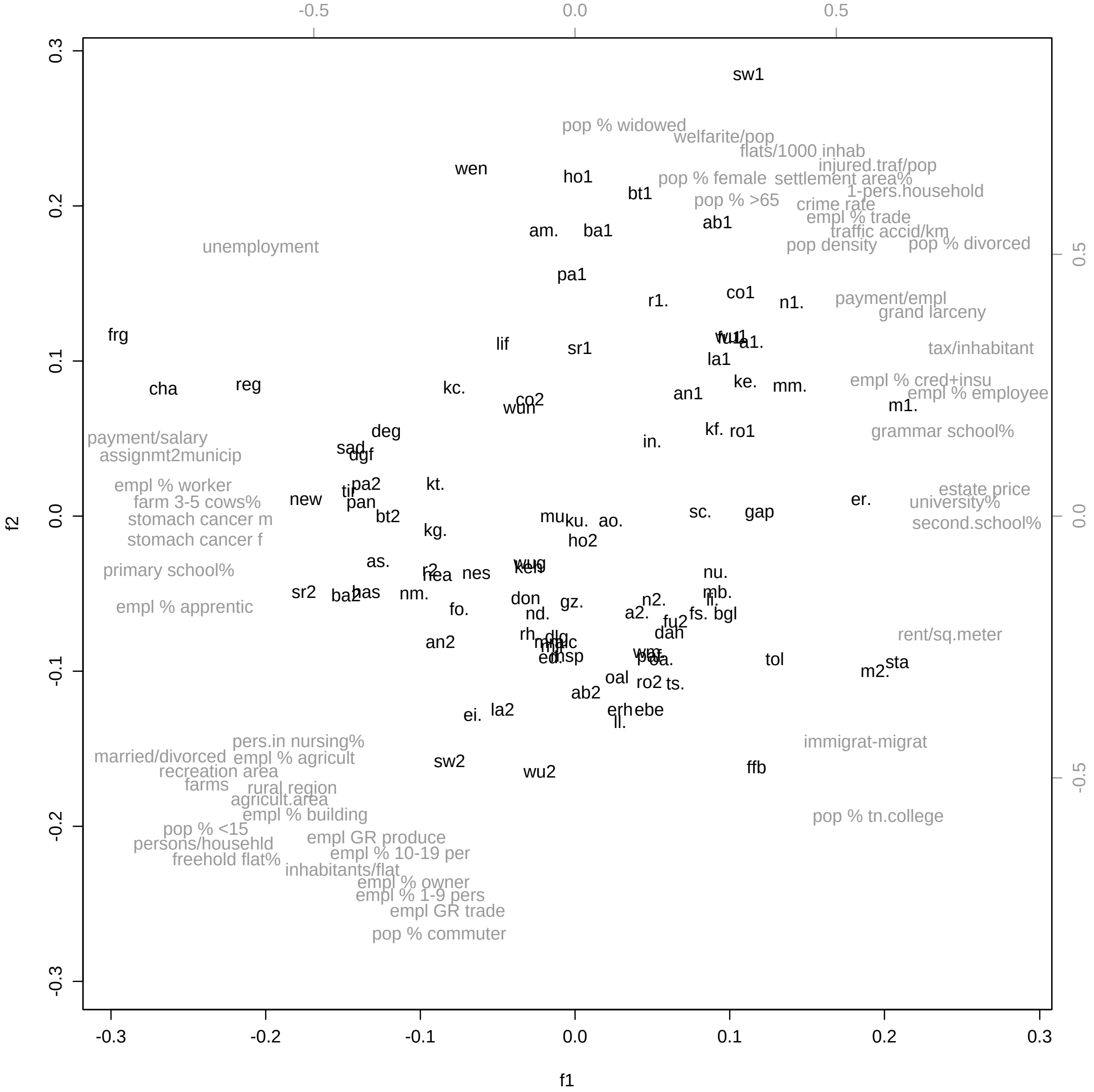

Figure 1: Plane with the central variables *stomach cancer* and *unemployment*

In the south-west of our diagram we have the rural regions (`rural region`) and recreation areas (`recreation area`) (both variables measured by inquiries), characterized by a high relation of married to divorced people (`married/divorced`), a large portion of persons in nursing (`pers.in nursing%`), a large portion of persons employed in the agriculture (`empl % agricult`), many farms (`farms`) and agricultural areas (`agricult.area`), a large portion of people employed in the building trade (`empl % building`), a large portion in children (`pop % <15`), a high relation of persons per household (`persons/ househld`), a large portion in freehold flats (`freehold flat%`), a high relation of inhabitants per flat (`inhabitants/flat`). Further we have some variables describing growth: growth rate in the producing trade (`empl GR produce`) and in the trade (`empl GR trade`), a large portion of employed people in companies up to 9 (`empl % 1-9 pers`) and from 10 to 19 employees (`empl % 10-19 per`), a large portion of business owners (`empl % owner`), and a large portion of commuters in the population (`pop % commuter`).

The opposite direction (north-east) contains variables typical for urban areas: grand larceny (`grand larceny`), a high relation of payment per employed people (`payment/ empl`), a large portion in divorced persons (`pop % divorced`), high population density (`pop density`), many accidents per kilometer in traffic (`traffic accid/km`), a large portion of employed people in trade (`empl % trade`), a high crime rate (`crime rate`), many one-person households (`1-pers.household`), a large portion in settlement area (`settlement area%`), many traffic accidents with personal injury (per 1000 inhabitants) (`injured.traf/pop`). Finally, we have some variables characterizing the rise in the ratio of old people to the total population: a large portion in people older than 65 years (`pop % >65`), a large portion in women (`pop % female`), a large number of flats per thousand inhabitants (`flats/1000 inhab`), many welfarite persons (`welfarite/pop`), and a large portion in widowed people (`pop % widowed`).

5 Summary

This method enables special views of a possibly large data set. The direction of the view is given, more or less, by selected variables. The selection of variables may be done by an interpretational point of view. It is also possible to select some variables of interest and, as shown in the example, additionally highly correlated variables. Since only the selected variables are analyzed by *PCA* or *FA*, the resulting subspace will describe these variables rather good. With a projection of the remaining variables into this subspace the relations to the variables of special interest can be shown. For interpretational purposes only variables which are close to the subspace should be projected. As a measure of closeness the *SMC* coefficient between variables and factors may be taken. A biplot representation of the subspace spanned by the factors, or better, two-dimensional subspaces spanned by all different pairs of factors, additionally shows the relations between variables and objects.

References

A. Basilevsky. *Statistical Factor Analysis and Related Methods: Theory and Applications*. Wiley & Sons, New York, 1994.

P. Filzmoser. *Principal Planes*. PhD thesis, Dept. of Statistics and Prob. Th., Vienna University of Technology, 1996. Unpublished.

K.R. Gabriel. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467, 1971.

H.H. Harman. *Modern Factor Analysis*. The University of Chicago Press, Chicago and London, 2nd edition, 1967.

D.N. Lawley and A.E. Maxwell. *Factor Analysis as a Statistical Method*. Butterworths, London, 1971.

K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis*. Acad. Press, London, 1979.

STUDIA and ALBTUM. External service of the rural agriculture in Bavaria. Technical report, STUDIA - Research Group for International Analyses, Schlierbach; ALBTUM - Professorship for Applied Agricultural Business Economics, Freising-Weihenstephan, Munich University of Technology, 1993. (in German).

Author’s address:

Dipl.-Ing. Dr. Peter Filzmoser
Institute of Statistics and Probability Theory
Vienna University of Technology
Wiedner Hauptstr. 8-10
A-1040 Vienna
Austria

Tel. +43 1 58801 / 5431
Fax +43 1 586-80-93
Elec. Mail: P.Filzmoser@tuwien.ac.at