R practice: Factor analysis

Minato Nakazawa (minato-nakazawa@umin.net)

27 June 2011

This material is just a rough draft to be read carefully with attention. Any suggestions and comments are welcome.

1 References

http://www.psy.ed.ac.uk/people/tbates/lectures/methodology/, which is provided by Prof. Timothy Bates, psychologist in the University of Edinburgh, is very helpful.

http://aoki2.si.gunma-u.ac.jp/lecture/PFA/pfa6.html[*1] is provided by Prof. Shigenobu Aoki, which are very informative but unfortunately written in Japanese.

2 The purpose of factor analysis

Official explanation Find the hidden factors behind observed variables: The hidden factors cannot measured directly, but should be "natural groupings"[*2] of observed variables.

Practical explanation Reduction of the number of variables intercorrelated: In this meaning, it resembles the principal component analysis.

3 Basic usage of factor analysis

Input Large numeric matrices, usually **more than 300 cases** with many variables. **Subjects/Variables ratio** should usually range **between 2:1 to 10:1**. Variables should obey normal distribution. Outliers should be omitted. Any variables uncorrelated with any other variables should be omitted. No variables correlated 1.0 with each other can be included: Remove one of each pair or take sum of them if appropriate.

Output (1) Factor loadings, which mean the correlation of each variable with the underlying factor (for that purpose, various rotations will be applied[*3]), (2) Factor scores, the summation of subject's responses $\times$ factor loadings, which mean the extent of each subject being explained by that factor.

Rotation There are two kinds of rotations: Orthogonal rotations keep independence among factors, but oblique rotations allow correlations among factors. If the factors may theoretically allow interdependence, the latter should be considered. The former includes the varimax rotation, which is most common and simple: to maximize squared column variance). The latter includes promax and oblimin rotations.

Tools Screeplot, Bartlett's sphericity test, Kaise-Meyer-Olkin's sampling adequacy criteria, and Parallel analysis are useful. After the successful factor extraction, Cronbach's α can be calculated to check whether the variables in each factor consist a uni-directional additive score or not (usually Cronbach's α must be more than 0.7 to consist a reliable scale).

When we interpret the extracted factors, adequate names (meaning) of factors are necessary. Well-defined factor should have at least three high-loading variables (if only one or two high loading(s), factors may be overextracted or multicolinearity may exist).

[*1] http://aoki2.si.gunma-u.ac.jp/R/kmo.html and http://aoki2.si.gunma-u.ac.jp/R/Bartlett.sphericity.test.html provide function definitions for KMO, MSA and Bartlett's sphericity test.

[*2] Subsets of variables that correlate strongly with each other and weakly with other variables in the dataset. Found factors should correspond to underlying "dimensions", which can be theoretically interpreted.

[*3] The initial factor loadings were calculated to maximize the loadings on the first factor, so that most items have large loadings on more than one factor, where the interpretation of the factor is difficult. Adequate rotation may solve this problem.

4 Basic model

Let's consider 10 variables $X_1, X_2, ..., X_{10}$ for 300 individuals. If we can assume 2 latent (hidden) factors F_1 and F_2 behind these 10 variables, each variable can be explained by these factors as follows.

$$X_1 = \beta_{1\cdot1}F_1 + \beta_{2\cdot1}F_2 + \epsilon_1$$

$$X_2 = \beta_{1\cdot2}F_1 + \beta_{2\cdot2}F_2 + \epsilon_2$$

$$\vdots$$

$$X_{10} = \beta_{1\cdot10}F_1 + \beta_{2\cdot10}F_2 + \epsilon_{10}$$

Here β means the correlation of each variable with the underlying factor: we call it **"Factor loadings"**. The ϵ means error variance, which is, in other word, the **uniqueness**, which cannot be explained by extracted factors. However, F_1 and F_2 are not measured. So we must estimate them by various method (principal axis method, minimum residual method, maximum likelihood method, and so on.) with iteration[*4].

Before rotation, F_1 and F_2 are assumed to be independent. If we denote n_{th} (n in [1,300]) individual's data of i_{th} (i in [1,10]) variable as $X_i(n)$, the "Factor scores" (here $FS_1(n)$ and $FS_2(n)$) can be obtained as follows (this is the simplest method. There are some other methods to estimate factor scores). Used variables for calculation are limited to have the absolute of β being large enough (usually more than 0.3, 0.4 or 0.5).

$$FS_1(n) = \sum_{i=1}^{10} \beta_{1\cdot i} X_i(n)$$

$$FS_2(n) = \sum_{i=1}^{10} \beta_{2\cdot i} X_i(n)$$

5 How many number of factors should be extracted?

There are some criteria, but no 100% foolproof statistical test exists.

- Drawing screeplot: Connect eigenvalues (as representing variances explained by each factor, so that sometimes sums of squared factor loadings are used instead) for many possible factors from maximum to minimum. The adequate number of factors is before the sudden downward inflextion of the plot.
- Parallel analysis: Compare actual screeplot with the possible screeplot based on randomly resampled data. The adequate number of factors is at the crossing point of the two plots.
- Eigenvalues > 1: Eigenvalues sum to the number of items, so an eigenvalue more than 1 is more informative than a single average item.

6 Checking adequacy of factor analysis

There are some method to check the adequacy of the factor analysis.

- Criteria of sample size adequacy: sample size 50 is very poor, 100 poor, 200 fair, 300 good, 500 very good, and more than 1,000 excellent (Comfrey and Lee, 1992, p.217).

[*4] In principal component analysis, each components can be formulated as the linear function of measured variables, so that it doesn't need iterative estimation.

- Kaiser-Meyer-Olkin's sampling adequacy criteria (usually abbreviated as KMO) with MSA (individual measures of sampling adequacy for each item): Tests whether there are a significant number of factors in the dataset: Technically, tests the ratio of item-correlations to partial item correlations. If the partials are similar to the raw correlations, it means the item doesn't share much variance with other items. The range of KMO is from 0.0 to 1.0 and desired values are > 0.5[*5]. Variables with MSA being below 0.5 indicate that item does not belong to a group and may be removed form the factor analysis.

Prof. Shigenobu Aoki provides the following function to calculate KMO and MSA at his web page:

```
kmo <- function(x)
{
        x <- subset(x, complete.cases(x))       # Omit missing values
        r <- cor(x)                             # Correlation matrix
        r2 <- r^2                               # Squared correlation coefficients
        i <- solve(r)                           # Inverse matrix of correlation matrix
        d <- diag(i)                            # Diagonal elements of inverse matrix
        p2 <- (-i/sqrt(outer(d, d)))^2          # Squared partial correlation coefficients
        diag(r2) <- diag(p2) <- 0               # Delete diagonal elements
        KMO <- sum(r2)/(sum(r2)+sum(p2))
        MSA <- colSums(r2)/(colSums(r2)+colSums(p2))
        return(list(KMO=KMO, MSA=MSA))
}
```

- Bartlett's sphericity test: Tests the hypothesis that correlations between variables are greater than would be expected by chance: Technically, tests if the matrix is an identity matrix. The p-value should be significant: *i.e.*, the null hypothesis that all off-diagonal correlations are zero is falsified.

Prof. Shigenobu Aoki provides the following function to conduct Bartlett's sphericity test at his web page:

```
Bartlett.sphericity.test <- function(x)
{
       method <- "Bartlett's test of sphericity"
       data.name <- deparse(substitute(x))
       x <- subset(x, complete.cases(x)) # Omit missing values
       n <- nrow(x)
       p <- ncol(x)
       chisq <- (1-n+(2*p+5)/6)*log(det(cor(x)))
       df <- p*(p-1)/2
       p.value <- pchisq(chisq, df, lower.tail=FALSE)
       names(chisq) <- "X-squared"
       names(df) <- "df"
       return(structure(list(statistic=chisq, parameter=df, p.value=p.value,
              method=method, data.name=data.name), class="htest"))
}
```

7 Functions to conduct factor analysis in R

factanal This function is included in standard installation. It uses maximum likelihood estimation (mle) to find the factor loadings. The number of factors to be extracted must be explicitly specified. Varimax and promax rotations are possible. Input data may be a matrix or a dataframe.

paf This function is included in **rela** package. It uses principal axis method to find the factor loadings. The

[*5] According to the criteria suggested by Kaiser (1974), less than 0.5 is unacceptable, [0.5, 0.6) is miserable, [0.6, 0.7) is mediocre, [0.7, 0.8) is middling, [0.8, 0.9) is meritorious, [0.9, 1.0) is marvelous.

adequate number of factors will be automatically determined by the criteria of eigenvalues (you can specify its criterion by `eigencrit=` option: default is 1). KMO and MSA are automatically calculated. Rotation is not provided. Input data must be a matrix.

fa This function is included in **psych** package. The `fm=` option can specify the method of estimation (`"minres"` for minimum residual, `"ml"` for maximum likelihood estimate, and `"pa"` for principal axis method). The number of extracted factors must be specified by `nfactors=` option. Various rotation methods can be specified by `rotate=` option (`"none"`, `"varimax"`, `"quartimax"`, `"bentlerT"`, `"geominT"`, `"oblimin"`, `"simplimax"`, `"bentlerQ"`, `"geominQ"`, and `"cluster"` will be possible).

alpha This function is included in **psych** package. This calculates Cronbach's α.

cortest.bartlett This function is included in **psych** package. This conducts Bartlett's sphericity test.

fa.parallel This function is included in **psych** package. Return the adequate number of extracted factors as `$nfact`.

8 Example 1

Let's analyze the variable `p1-p40` in the `factorexdata05.txt`, which is converted from Prof. Timothy Bates' SPSS data[*6]. Prof. Bates provides the pdf documents for undergraduate students[*7].

The easiest way is the following. Number of factors can be automatically determined. Factor loadings are saved as `res$Factor.Loadings`.

```
library(foreign)
y <- read.spss("http://www.subjectpool.com/ed_teach/y3method/factorexdata05.sav")
x <- as.data.frame(y)
for (i in 1:length(x)) { x[,i] <- ifelse(x[,i]==999,NA,x[,i]) }
# The data \verb!x! consists of 538 cases with 102 variables.
# it can be saved as "factorexdata05.txt" by the following line
# write.table(x,"factorexdata05.txt",quote=FALSE,sep="\t",row.names=FALSE)
# if so, the data can be read by:
# x <- read.delim("factorexdata05.txt")
Ps <- x[,4:43] # Extract variables p1-p40
Ps <- subset(Ps, complete.cases(Ps)) # Omit missings (511 cases remain)
library(rela)
res <- paf(as.matrix(Ps))
summary(res) # Automatically calculate KMO with MSA, determine the number of factors,
             # calculate chi-square of Bartlett's sphericity test, communalities and
             # factor loadings.  Communalities are 1 minus uniquenesses.
barplot(res$Eigenvalues[,1]) # First column of eigenvalues.
resv <- varimax(res$Factor.Loadings) # Varimax rotation is possible later.
print(resv)
barplot(sort(colSums(loadings(resv)^2),decreasing=TRUE)) # screeplot using rotated SS loadings.
scores <- as.matrix(Ps) %*% as.matrix(resv$loadings) # Get factor scores in a simple manner.
library(psych)
cortest.bartlett(Ps) # Bartlett's sphericity test.
res2 <- fa.parallel(Ps)
res3 <- fa(Ps, fm="minres", nfactors=8, rotate="oblimin")
print(res3) # Factor loadings as $loadings
```

[*6] `http://www.subjectpool.com/ed_teach/y3method/factorexdata05.sav`

[*7] `http://www.subjectpool.com/ed_teach/y3method/factorex05.pdf` and `http://www.subjectpool.com/ed_teach/y3method/fa.pdf`