Exploratory factor analysis

Jose M Sallan
jose.maria.sallan@upc.edu

Quantitative research methods

April 17, 2018

Outline

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Superior d'Enginyeries Industrial, Aeroespacial i Audiovisual de Terrassa

The aim of factor analysis is to explain the variability of a set of random, correlated observable variables in terms of a lower number of unobservable, latent variables called **factors**

Exploratory and confirmatory factor analysis

Exploratory factor analysis (EFA)

- No *a priori* assumption is made about the number and nature of factors
- The obtained factors must be interpreted *a posteriori*, examining which observable variables correlate with each factor

Confirmatory factor analysis (CFA)

- Testing a model where there is an *a priori* assumption about which variables are associated with each factor
- CFA is performed using **structural equation modelling**

Component and common factor models

Two possible mathematical approaches to EFA:

Component model

- variables observed without error
- estimated through the **principal components** method

Common factor models

- variables include measurement error
- several methods of estimation: **principal axis**, **maximum likelihood**

Example: student's scores

Correlation matrix of student's scores on six subjects:

	Gaelic	English	History	Arithmetic	Algebra	Geometry
Gaelic	1.000	0.439	0.410	0.288	0.329	0.248
English	0.439	1.000	0.351	0.354	0.320	0.329
History	0.410	0.351	1.000	0.164	0.190	0.181
Arithmetic	0.288	0.354	0.164	1.000	0.595	0.470
Algebra	0.329	0.320	0.190	0.595	1.000	0.464
Geometry	0.248	0.329	0.181	0.470	0.464	1.000

How would you group subjects? How are the correlations among subjects in the same group?

Example: student's scores

We can analyze a 2-component model using the principal function of the psych package in R:

```
> library(psych)
> pr.scores <- principal(r=scores.cor, nfactors=2,
+                                  rotate="none")
```

Example: student's scores

And we can also obtain a 2-common factor model using the `fa` function of the psych package (with `fm="pa"` for principal axis and `fm="ml"` for maximum likelihood)

```
> library(psych)
> pa.scores <- fa(r=scores.cor, nfactors=2, fm="pa",
+                     rotate="none")
> ml.scores <- fa(r=scores.cor, nfactors=2, fm="ml",
+                 rotate="none")
```

Factor loadings

- The factor loadings are the correlation between a variable and a factor
- If a variable i has a large loading on factor j means that a large amount of the variability of i can be explained by j
- Hopefully each variable has a high loading in a single factor

Factor loadings

The loadings of the principal components:

```
> pr.scores$loadings

Loadings:
           PC1    PC2
Gaelic      0.658  0.449
English     0.688  0.290
History     0.517  0.637
Arithmetic  0.738 -0.413
Algebra     0.744 -0.375
Geometry    0.678 -0.355

                 PC1   PC2
SS loadings    2.733 1.130
Proportion Var 0.455 0.188
Cumulative Var 0.455 0.644
```

Rotation

The values of factor loadings are not unique: any **rotation** of the loadings is also a valid solution

There are several rotation methods that try to attach each variable to a single factor:

- Orthogonal (**varimax**, **quartimax**): keep factors uncorrelated (independent)
- Non-orthogonal (**oblimin**)

Default rotation method in R commands is oblimin

Example: student's scores

Defining rotations

Defining rotations...

```
> pr.scores.varimax <- principal(r=scores.cor, nfactors=2,
+                                rotate="varimax")
> pa.scores.oblimin <- fa(r=scores.cor, nfactors=2,
+                             fm="pa")
> ml.scores.quartimax <- fa(r=scores.cor, nfactors=2, fm="ml",
+                 rotate="quartimax")
```

Example: student's scores

Rotated solution

```
> pr.scores.varimax$loadings

Loadings:
           RC1   RC2
Gaelic     0.223 0.765
English    0.348 0.661
History          0.821
Arithmetic 0.833 0.150
Algebra    0.813 0.182
Geometry   0.749 0.156

                 RC1   RC2
SS loadings    2.087 1.776
Proportion Var 0.348 0.296
Cumulative Var 0.348 0.644
```

Factor loadings of rotated solution


```
> print(pr.scores.varimax$loadings, cutoff=0.4)

Loadings:
           RC1    RC2
Gaelic             0.765
English            0.661
History            0.821
Arithmetic  0.833
Algebra     0.813
Geometry    0.749

                  RC1   RC2
SS loadings     2.087 1.776
Proportion Var 0.348 0.296
Cumulative Var 0.348 0.644
```

Example: regulatory focus scales

150 observations of two 6-items scales measuring promotion and prevention focus

- **Promotion focus:** processes that support completion of tasks by strategically approaching means necessary to accomplish the task
- **Prevention focus:** processes that support completion of tasks by strategically avoiding those things that may deter successful task execution

Example: regulatory focus scales

Picking data and computing correlations:

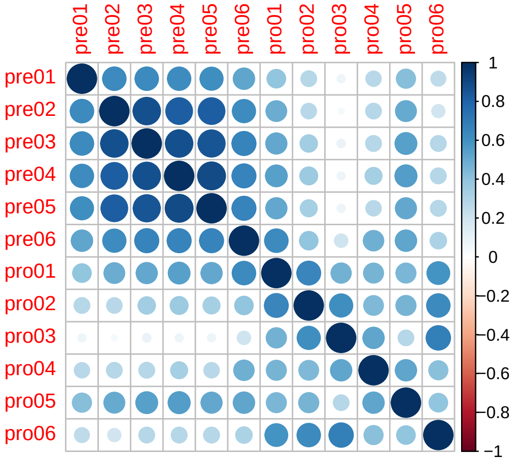
```
> data <- read.csv("datascale.csv")
> vars <- names(data)
> focus <- data[
+   ,which(grepl("pre", vars) | grepl("pro", vars))]
> cor.focus <- cor(focus)
```

Example: regulatory focus scales

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Superior d'Enginyeries Industrial, Aeroespacial i Audiovisual de Terrassa

Correlogram of variables

```
> library(corrplot)
> corrplot(cor.focus, method="circle")
```



Preliminary analysis

Testing existence of correlations

Prior to performing EFA, it has to be checked whether the variables are correlated:

- **Kaiser-Meyer-Olkin sample adequacy test:** test whether partial correlations are large enough (should be larger than 0.5)
- **Bartlett's sphericity test:** population correlation matrix equal to identity as null hypothesis (must be rejected)

Example: regulatory focus scales

Preliminary tests

```
> library(psych)
> KMO(cor.focus)

Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = cor.focus)
Overall MSA =  0.89
MSA for each item =
pre01 pre02 pre03 pre04 pre05 pre06 pro01 pro02 pro03 pro04 pro05 pro06
 0.97  0.90  0.90  0.90  0.90  0.91  0.88  0.89  0.75  0.83  0.93  0.83

> cortest.bartlett(cor.focus, n=150)

$chisq
[1] 1429.374

$p.value
[1] 3.522197e-255

$df
[1] 66
```

Preliminary analysis

Number of factors to extract

How many factors do we extract?

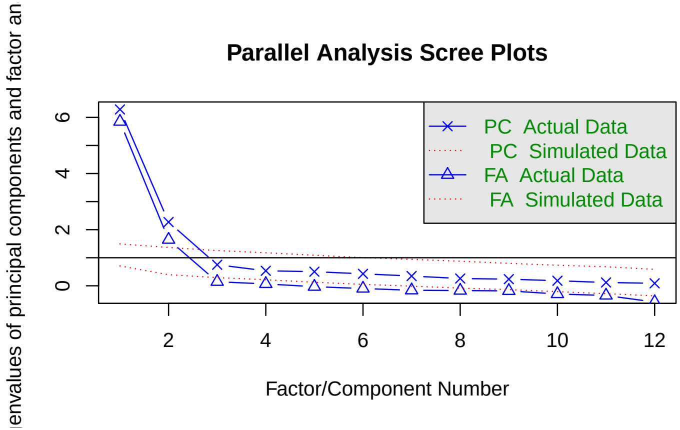
- Theoretical considerations (two factors for the RF example)
- For the component model: number of eigenvalues larger than one
- Proportion of overall variance explained (around 0.85)
- Scree analysis: `fa.parallel` function

Example: regulatory focus scales

Number of factors to extract

```
> fa.parallel(cor.focus, n.obs=150)

Parallel analysis suggests that the number of factors =  2  and the number of components =  2
```



Example: regulatory focus scales

Model extraction

As scales are variables with measurement error, it is preferable to use a common factor model: principal axis with two factors:

```
> pa.data.oblimin <- fa(r=cor.focus, nfactors=2, fm="pa")
```

Example: regulatory focus scales

Model results

```
> pa.data.oblimin

Factor Analysis using method =  pa
Call: fa(r = cor.focus, nfactors = 2, fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
        PA1   PA2   h2   u2 com
pre01  0.67  0.04 0.47 0.53 1.0
pre02  0.93 -0.08 0.82 0.18 1.0
pre03  0.94 -0.02 0.87 0.13 1.0
pre04  0.93  0.00 0.87 0.13 1.0
pre05  0.92 -0.02 0.84 0.16 1.0
pre06  0.66  0.23 0.59 0.41 1.2
pro01  0.40  0.58 0.64 0.36 1.8
pro02  0.14  0.74 0.63 0.37 1.1
pro03 -0.21  0.90 0.73 0.27 1.1
pro04  0.18  0.55 0.41 0.59 1.2
pro05  0.48  0.34 0.46 0.54 1.8
pro06  0.04  0.77 0.62 0.38 1.0

                       PA1  PA2
SS loadings           5.01 2.94
Proportion Var        0.42 0.24
Cumulative Var        0.42 0.66
Proportion Explained  0.63 0.37
Cumulative Proportion 0.63 1.00

 With factor correlations of
     PA1  PA2
PA1 1.00 0.33
PA2 0.33 1.00

Mean item complexity =  1.2
Test of the hypothesis that 2 factors are sufficient.
```

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Superior d'Enginyeries Industrial, Aeroespacial i Audiovisual de Terrassa

- **h2** (communalities): variability explained by FA
- **u2** (uniquenesses): variability not explained by FA
- **Proportion Var**: proportion of variance explained by factor
- **Cumulative Var**: cumulative explained variance

In the RF example, the correlogram has advanced a good deal of results obtained

Factor interpretation

- The meaning of factors has to do with the meaning of variables with high loading in factor
- Student's scores example: verbal vs numerical capabilities

Example: regulatory focus scales

Factors intepretation


```
> print(pa.data.oblimin$loadings, cutoff=0.4)

Loadings:
      PA1    PA2
pre01  0.673
pre02  0.926
pre03  0.937
pre04  0.934
pre05  0.922
pre06  0.665
pro01         0.576
pro02         0.739
pro03         0.901
pro04         0.553
pro05  0.481
pro06         0.772

                 PA1   PA2
SS loadings    4.840 2.769
Proportion Var 0.403 0.231
Cumulative Var 0.403 0.634
```

What is the intepretation of each factor?

Factor scores

Once performed the factor analysis, it is possible to obtain values of the factors for each observation.
For the RF example:

```
> pa.data.oblimin.scores <-
+   factor.scores(focus, pa.data.oblimin, method="Thurstone")
> scores <-
+   as.data.frame(pa.data.oblimin.scores$scores)
```

Example: regulatory focus scales

```
> colors <- rep("blue", 150)
> colors[which(data$sexo==1)] <- "red"
> plot(scores$PA1, scores$PA2, pch=19, col=colors, xlab="PA1 (prevention)", ylab="PA2 (promotion)",
+      cex.lab=0.7, cex.axis=0.7)
> abline(lm(scores$PA2 ~ scores$PA1))
```