

Factor analysis

Problem Statement

Biologists often encounter high-dimensional datasets from which they wish to extract features which are known to be represented within the data (e.g. clinicians may suspect several diseases given a list of clinical signs and symptoms in patients with respiratory distress. Factor analysis may be used to separate patients according to clinical signs)


Factor analysis (FA) is another dimensionality reduction method used to identify latent features (or factors) in a dataset from a set of variables found in the original dataset

FA works in a similar way to PCA, creating a linear combination of factors that represent similarity between variables

Problem Statement - exam example

Factor 1: mathematical ability

Factor 2: writing ability



	Arithmetic	Algebra	Geometry	Statistics	Creative writing	Literature	Spelling/Grammar
Student 1	45	22	67	58	76	87	93
Student 2	22	53	75	19	86	69	90
Student 3	90	94	75	85	32	46	55
Student 4	56	63	68	53	67	61	58
Student 5	44	52	35	64	77	82	76
Student 6	89	79	93	94	56	43	60

Example

The prostate data have **97** rows and **9** columns.

Columns (red = clinical continuous variables):

lcavol (log-transformed cancer volume)

lweight (log-transformed prostate weight)

lbph (log-transformed amount of benign prostate enlargement)

svi (seminal vesicle invasion)

lcp (log capsular penetration; amount of spread of cancer in outer walls of prostate)

gleason (Gleason score; grade of cancer cells)

pgg45 (percentage Gleason scores 4 or 5)

lpsa (log prostate specific antigen; level of PSA in blood)

age (patient age in years)

Example

We will calculate factors using clinical data recorded in the prostate dataset, each of which is a continuous variable, so that we can create fewer variables representing clinical markers of cancer progression.

We will create a new dataset only including the variables lcavol, lweight, lbph, lcp and lpsa. We know from our PCA analysis that two principal components explain an adequate amount of variation in the data which may help in our initial choice of factors.

Challenge 1

Use the 'factanal' function to identify the minimum number of factors necessary to explain most of the variation in the data

Output of FA

Loadings: values between -1 and 1 which represent contribution of variables to factors. Positive values suggest positive relationship between variable and factor and negative values suggest negative relationship between variable and factor

Communality: the degree to which variables explain the same variation in the data and is calculated for each variable by summing the (squared) loadings

Uniqueness: the amount of variation in the data uniquely explained by one variable. Uniqueness is calculated by subtracting the communality value from 1

$$\text{Variability in data } (\hat{\Sigma}) = \text{Communality} + \text{Uniqueness}$$

Challenge 2

Use the output from your factor analysis and the plots above to interpret the results of your analysis.

What variables are most important in explaining each factor? Do you think this makes sense biologically?
Discuss in groups.

Advantages/disadvantages of FA

Advantages:

- easy to implement with many softwares/packages available
- can take into account user knowledge on number of factors to include
- used in many different fields

Disadvantages:

- number of factors must be predefined by the user
- interpreting factors can be difficult
- factor analysis is not often used in some fields (e.g. genomics, medical sciences), although this is changing
- only continuous variables can be included in a basic factor analysis

Further reading

Gundogdu et al. (2019) Comparison of performances of Principal Component Analysis (PCA) and Factor Analysis (FA) methods on the identification of cancerous and healthy colon tissues. International Journal of Mass Spectrometry 445:116204

Kustra et al. (2006) A factor analysis model for functional genomics. BMC Bioinformatics 7:
doi:10.1186/1471-2105-7-21

Yong, A.G. & Pearce, S. (2013) A beginner's guide to factor analysis: focusing on exploratory factor analysis. Tutorials in Quantitative Methods for Psychology 9(2):79-94