Lab 12

Wiktor Soral

May 23rd 2017

Problem of too many dimensions

- Suppose you would like to compare cognitive capabilities among a group of 4 individuals. You have administered 4 different tests. Below are the results. Who has the highest cognitive capabilities?

	t1	t2	t3	t4
I1	5	4	6	5
I2	7	8	2	3
I3	1	2	8	9
I4	6	7	3	4

Problem of too many dimensions

Similar problems are frequently encountered in empirical science.

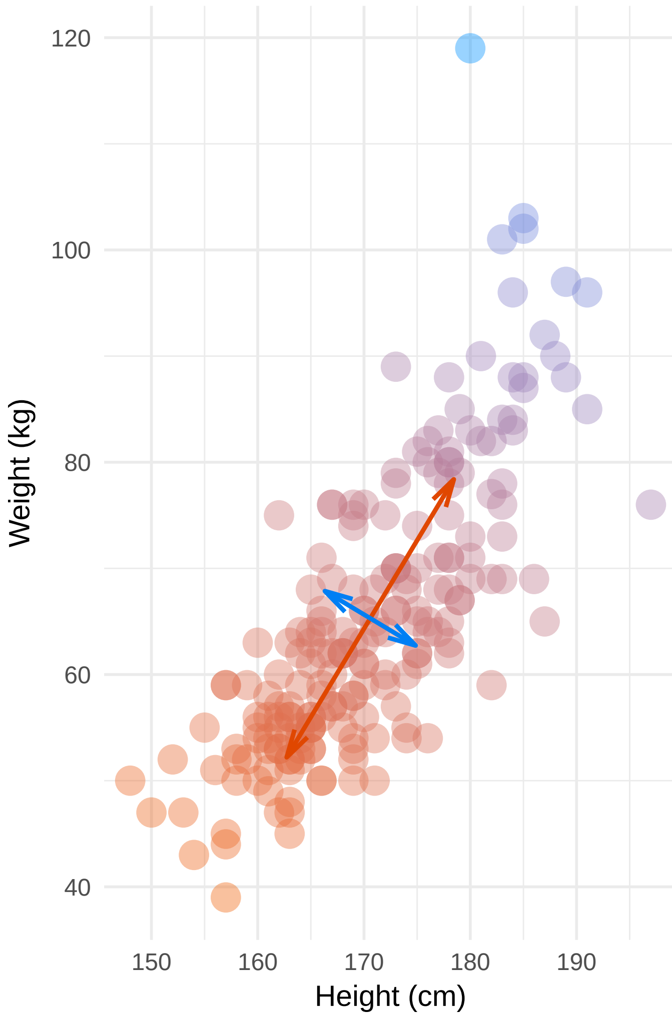
- You would like to assess a general level of prejudice having obtained prejudice towards 10 different minorities.
- You would like to obtain a score from a long and complex personality questionnaire.
- You would like to obtain scores measuring how other person's are perceived, having obtained evaluations on 30 different traits.

Each item from the questionnaire or each trait forms a separate dimension. Can you imagine a 30-dimensional space?

Reducing dimensionality

- To deal with the problem of multidimensionality we want to make a projection on a space with lower number of dimensions.
- In other words we want to find a small number of dimensions, that would explain most of initial variation.
- The ideal space would explain 100% of the initial variation with only 1 dimension.
- Usually the aim is to find space that would explain at least 50% of variation, with a few dimensions (2-4)

Reducing dimensionality

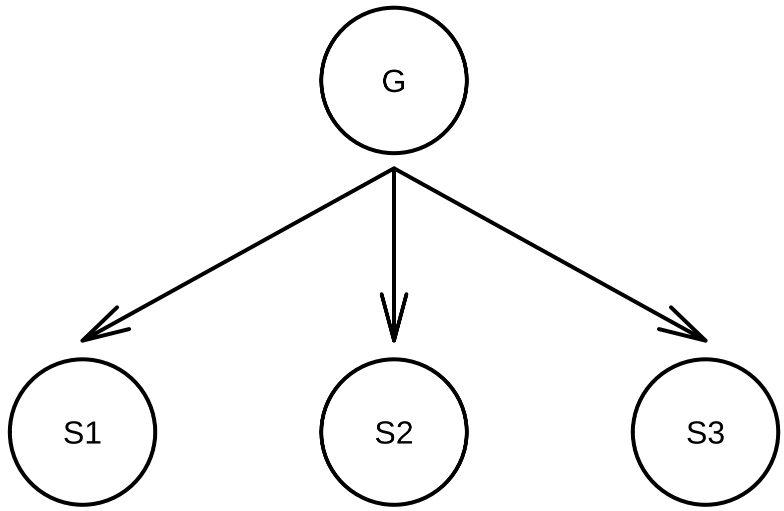


Reducing dimensionality

```
## Importance of components:
##                           Comp.1     Comp.2
## Standard deviation     15.2364029 4.96837833
## Proportion of Variance  0.9038877 0.09611229
## Cumulative Proportion   0.9038877 1.00000000
```

```
##
## Loadings:
##        Comp.1 Comp.2
## height  0.515 -0.857
## weight  0.857  0.515
##
##                Comp.1 Comp.2
## SS loadings       1.0    1.0
## Proportion Var    0.5    0.5
## Cumulative Var    0.5    1.0
```

Factor analysis

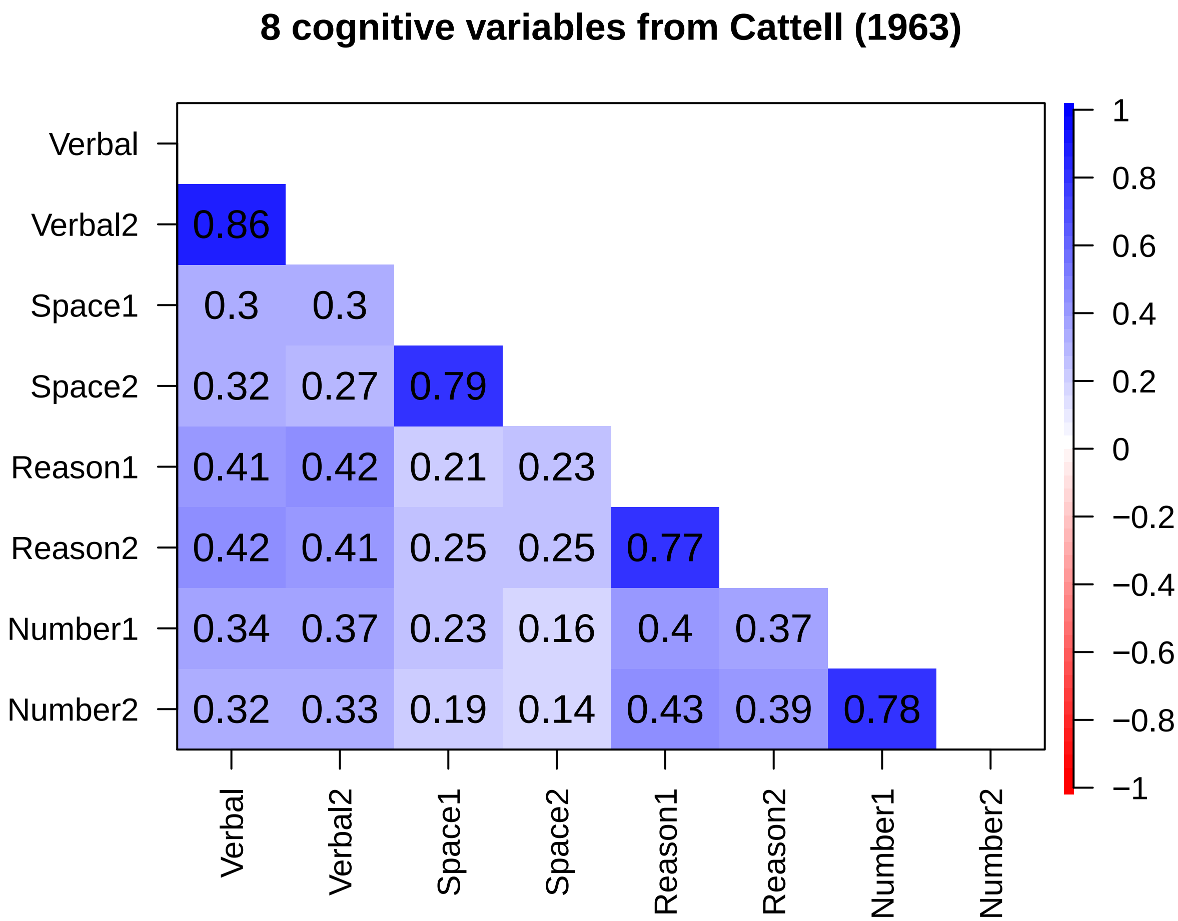


Factor analysis

The factor analysis model is $x = \Lambda f + \epsilon$

- x is p element vector of observable responses, e.g. reponses of an individual to questionnaire with p items
- f is k element vector of unobservable (latent) scores of an individual, e.g. Spearman's G factor
- Λ is $p \times k$ matrix of loadings, i.e. how much latent factor affects responsnes to each item of measurement
- ϵ is a p element vector of errors, i.e. amount of variation of observable responses not explained by latent factor

Factor scores



Factor analysis - requirements

- Large sample size - a common rule is to have at least 10-15 participants per variable
- Patterns in the correlation matrix should not be diffuesed - check Kaiser-Meyer-Olkin measure of sampling adequacy (KMO) - if greater than 0.7 it is good; should not be below 0.5
- Variables should correlate with each other - check whether correlation matrix differs from identity matrix with Bartlett test (p should be significant)
- Correlation matrix should not be singular - i.e. no multicollinearity - check determinant of R-matrix - should be greater than 0.00001
- Multivariate normality