Data Reduction Techniques

Jan Rovny

March 28, 2023

Introduction

- We live in a world which collects an increasing amount of quantitative data about almost anything
- We are faced with large datasets with an enormous amount of variables
- To make sense of the data in front of us – we need to seek structure
- There are different methods to reduce data, here are three:
  - Principal Component Analysis
  - Factor Analysis
  - Item Response Theory

Where does PCA and FA come from?

- PCA and FA was developed in the early 20th century by psychologists when studying human abilities.
- Charles Spearman noticed structure in human abilities: if a child writes well, she also reads well, counts well...
- The researchers thus discovered a 'factor' which collectively describes human ability. This is the 'General Intelligence Factor.'

What is PCA and FA for?

- FA shows patters of interrelationships between variables
- By clustering variables into groups, PCA or FA identifies
  - common *'components'* which summarize multiple variables
  - which variables are redundant
- By being able to combine variables that measure 'the same thing,' PCA or FA can be a solution in cases of multicollinearity.

Issues

- **Factor extraction** – different ways of simplifying our data:
  - Principal Component Analysis
  - Factor Analysis
- **Factor retention** – how many factors do we keep to meaningfully simplify our data
  - Kaiser versus Screeplot
- **Factor rotation** – how do we manipulate our factors to be best able to interpret them
  - Varimax, Promax, Oblimin...

Principle Component Analysis - PCA

- PCA is the simplest method of data reduction:
  - Imagine we have two variables, what is the easiest way to simplify them into one component?
  - Draw a line of best fit between them (this is the 'regression line').
  - If we define a variable depicted by this line, we have simplified our two variables into one 'component.'
- If we have more variables, we simply continue to fit additional lines until our factors account for all the variance in our original variables.
  - How many components are necessary to capture **all** the variance of K variables?
- PCA is thus a very simple data reduction technique, computed using all the variance of the items, without regard for any underlying data structure.

See Demonstration in R

Terminology

- **Eigenvalue** describes the variance ‘unique’ to the particular factor.
- Notice that the sum of all eigenvalues in PCA equals the number of initial items (21)
- **Proportion** translates the eigenvalues into a proportional measure of variance explained by each factor. Factor 1 thus explains over 50% of all variance.
- The **Cumulative** measure simply sums the proportional variance measures, showing that the first 3 factors explain 87% of all variance. All 21 factors (necessarily) explain 100%.

Component Retention

- So, now we have information about data structure, but how many components should we consider ‘sufficient’ to simplify our data?
- Kaiser Criterion
  - Keep all components whose eigenvalue > 1
  - This means we retain all components that capture more variance than a single item
  - This is criticized by the literature as inaccurate.
- Scree Test
  - Jump off a cliff and look for the elbow! (A little geological excursion)
  - Retain components above (and excluding) the elbow.
  - This is criticized as subjective...

Component Interpretation

- We retain 3 components. Now, what do they mean?
- Components have no statistically determined meaning, they must be interpreted from component loadings.

Table of Loadings

	PC1	PC2	PC3	PC4	PC5
eu_position	-0.51	0.62	0.13	0.34	0.06
immigrate_policy	0.95	-0.04	0.02	0.05	0
multiculturalism	0.95	-0.06	-0.02	0.03	0
redistribution	0.62	0.71	0.13	-0.17	0.04
environment	0.87	0.15	-0.06	0.01	-0.11
spendvtax	0.66	0.66	0.11	-0.21	0.04
deregulation	0.53	0.79	0.14	-0.19	0.02
econ_interven	0.49	0.78	0.21	-0.23	0.06
civlib_laworder	0.94	-0.08	-0.1	0.03	-0.02
sociallifestyle	0.9	-0.18	0	0.2	0.13
religious_principles	0.79	0.02	-0.13	0.26	0.17
ethnic_minorities	0.92	-0.08	-0.03	-0.02	-0.17
nationalism	0.93	-0.23	-0.12	0.03	0.07
urban_rural	0.63	-0.21	-0.39	0.13	0.46
protectionism	0.26	-0.88	-0.11	0.06	0.14
regions	0.36	-0.11	-0.03	0.22	-0.84
russian_interference	0.2	0.07	0.58	0.62	0.01
anti_islam_rhetoric	0.79	-0.26	0.16	-0.02	-0.15
people_vs_elite	0.09	-0.64	0.37	-0.52	0.02
antielite_salience	0.36	-0.69	0.44	-0.29	-0.04
corrupt_salience	-0.04	-0.24	0.79	0.19	0.23
EV	9.617	4.513	1.644	1.234	1.122

Factor Rotation

- The metric on which our component loadings are measured is, however, arbitrary.
- It is similar to the metric used for measuring Earth's longitude – 0 ° at Greenwich is an arbitrary outcome of British naval dominance...
- We can thus rotate the reference axes to increase the interpretability of our components. Here is an example from Abdi:

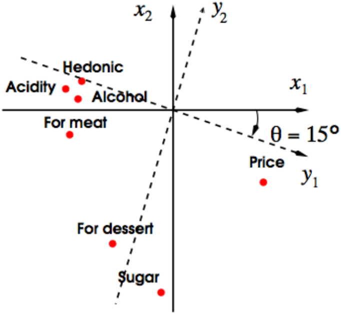

Figure 3: PCA: The loading of the seven variables for the wine example showing the original axes and the new (rotated) axes derived from VARIMAX.

Types of Component Rotation

- There are two classes of component rotation:
  - **Orthogonal** rotation keeps the axes independent on each other. The angle between the axes is 90 °. This is the case in the example above.
  - **Oblique** rotation relaxes the independence of axes, allowing the angle between the axes to vary.
- Both classes of rotation make intuitive sense:
  - Keeping axes orthogonal gives us a rational metric in which we can chart reality. Maps, after all, depict the world as flat with orthogonal latitude and longitude.
  - The assumption that our components are correlated (i.e. are not orthogonal, but oblique) is very reasonable. The world is not flat and latitude and longitude are orthogonal only on the equator!

Factor Rotation

- In reality, rotation slightly improves interpretation.

Varimax Rotated Component Loadings

	RC1	RC2	RC4	RC3
eu_position	-0.53	0.22	-0.6	0.28
immigrate_policy	0.87	0.3	0.16	0.08
multiculturalism	0.88	0.29	0.17	0.04
redistribution	0.34	0.9	-0.11	0.02
environment	0.75	0.42	0.04	-0.03
spendvtax	0.39	0.88	-0.06	-0.02
deregulation	0.23	0.94	-0.16	0.01
econ_interven	0.18	0.95	-0.1	0.04
civlib_laworder	0.89	0.24	0.14	-0.03
sociallifestyle	0.92	0.11	0.12	0.17
religious_principles	0.84	0.17	-0.12	0.1
ethnic_minorities	0.82	0.28	0.22	-0.03
nationalism	0.93	0.11	0.21	-0.04
urban_rural	0.82	-0.1	-0.07	-0.15
protectionism	0.48	-0.67	0.44	0
regions	0.26	-0.04	0	0
russian_interference	0.17	0.06	-0.14	0.83
anti_islam_rhetoric	0.69	0.14	0.38	0.13
people_vs_elite	0.01	-0.17	0.89	0.02
antielite_salience	0.29	-0.18	0.84	0.21
corrupt_salience	-0.1	-0.06	0.35	0.78

Principal Factor Analysis - PF

- An alternative to PCA is **Principal Factor Analysis PF**.
- It differs from PCA by assuming that the individual items measure some underlying 'latent' variables.
- Consequently, each item measures: 1) common underlying concept, and 2) some unique issue.
- PF uses only the variability in an item that it has in common with the other items to find the factor solution.
- PF thus assumes there is some underlying structure and uses only the common variance, while PCA makes no assumptions about the data and uses all the variance of the items.
- In practice, PF tends to produce fewer factors with eigenvalue > 1.

PF - An Example

	PA1	PA2	PA3	PA4
eu_position	-0.51	0.62	0.05	0.42
immigrate_policy	0.95	-0.04	0.01	0.07
multiculturalism	0.95	-0.06	-0.02	0.04
redistribution	0.62	0.71	0.16	-0.13
environment	0.86	0.14	-0.05	-0.02
spendvtax	0.66	0.66	0.14	-0.17
deregulation	0.53	0.79	0.17	-0.14
econ_interven	0.49	0.78	0.24	-0.15
civlib_laworder	0.94	-0.08	-0.1	0.02
sociallifestyle	0.91	-0.18	-0.05	0.26
religious_principles	0.79	0.01	-0.18	0.27
ethnic_minorities	0.92	-0.08	-0.01	-0.05
nationalism	0.93	-0.23	-0.13	0.01
urban_rural	0.63	-0.21	-0.43	0.09
protectionism	0.25	-0.86	-0.11	0.04
regions	0.35	-0.1	-0.03	0.04
russian_interference	0.19	0.06	0.34	0.5
anti_islam_rhetoric	0.78	-0.25	0.17	-0.01
people_vs_elite	0.09	-0.61	0.44	-0.35
antielite_salience	0.36	-0.68	0.51	-0.17
corrupt_salience	-0.04	-0.22	0.65	0.4
EV	9.542	4.429	1.420	0.988

PCA or PF?

A logical question is thus when to use PCA and when PF?

- Use PCA:
  - When we want to see the general structure of the data.
  - When we want to reduce our data to independent sets of items or variables.
- Use PF
  - When we want to identify latent underlying variables.
  - When we want to capture these latent variables for further statistical uses.
  - Since we often expect to see some latent variables, the literature leans in favour of PF.

Conclusion

- Factor analysis is a useful method for learning about our data.
- Factor analysis allows us to understand the underlying structure of our data, see its relationships and redundancies.
- Factor analysis allows us to find underlying latent variables which can be extracted and used in further analysis.
- This can be a solution to multicollinearity since collinear variables may measure the same thing, which can be captured and replaced in the analysis by a single factor.
- **Warning**: factor analysis is an *inductive* procedure. It relies solely on the data fed to it. **There is no theory, no hypotheses and no inference. It is just the data that speaks!**