Principal components analysis

Multivariate methods

Multivariate methods are used to analyse multiple outcome variables together, in comparison with all previous methods where there was only one outcome variable. They are used in general to try to reduce a complex dataset to a simpler one which is easier to interpret and understand.

What is principal components analysis?

This method is used to reduce a dataset with many inter-correlated variables to a smaller set of uncorrelated variables which explain the overall variability almost as well. It is sometimes described as '**reducing the dimensionality of a dataset**'. The derived smaller set of variables is then used in later analyses in place of the original larger set.

How principal components analysis works

- The method gives a set of **principal components** (**PCs**), each of which is a linear combination of all the original variables
- If there are *n* variables in total then a maximum of *n* PCs can be computed
- Each PC explains a proportion of the total variability
- The first PC is the one that explains the maximum amount of the variance and the second PC explains the next greatest amount and so on

Principal component equations

The following equations show how principal components analysis works mathematically and how the principal components are related to the original set of variables. Assuming that the original variables are:

$x_1, x_2, x_3 \ldots x_p$ the method produces p principal components $y_1, y_2, y_3 \ldots y_p$, which are defined as follows:

$$y_1 = b_{11}x_1 + b_{12}x_2 + \ldots + b_{1p}x_p$$

$$y_2 = b_{21}x_1 + b_{22}x_2 + \ldots + b_{2p}x_p$$

$$y_p = b_{p1}x_1 + b_{p2}x_2 + \ldots + b_{pp}x_p$$

where b_{11}, b_{12} etc. are coefficients.

Practicalities

- It is common practice to include enough PCs to explain at least 80% of the total variability and this often needs only two or three
- Principal components analysis provides a single value for each PC for each subject and therefore each PC is a new variable
- These are then used in further analyses in the same way as other variables are analysed

Interpreting principal components

Specific principal components sometimes usefully represent a particular overarching theme, where several of the original variables contribute to the theme. The example (📖 Principal components analysis: example, p. 442) illustrates this.

Principal components analysis: example

Example

Researchers wished to determine the important features of six lung function tests in 458 coalminers.[1] They used principal components analysis and reduced the six tests to three meaningful respiratory components. The results are summarized in Table 12.12.

Table 12.12 Coefficients for the first four principal components with six lung function variables

Component	1st	2nd	3rd	4th
FEV_1	–0.46	0.18	0.23	–0.26
FVC	–0.38	0.58	0.40	–0.22
FEV_1/FVC	–0.38	–0.57	–0.24	–0.52
$Vmax_{50}$	–0.44	–0.32	0.12	0.05
$Vmax_{25}$	–0.43	–0.21	0.17	0.77
TLCO	–0.35	0.41	–0.83	0.14
% variability	74%	15%	7%	3%

FEV_1, forced expiratory volume in 1 second; FVC, forced vital capacity; TLCO, transfer factor of the lung for carbon monoxide.

- The analysis produces six PCs but the four shown here explain virtually all of the overall variability (99%) in the six lung function measures
- The first principal component is:
  $-0.46 \times FEV_1 - 0.38 \times FVC - 0.38 \times FEV_1/FVC - 0.44 \times Vmax50 - 0.43 \times Vmax_{25} - 0.35 \times TLCO$
- The largest coefficients for the first PC were for forced expiratory volume in 1 second (FEV_1), $Vmax_{50}$, and $Vmax_{25}$, which measure the capacity of the lungs, and so the authors concluded that the first PC mainly represented **lung size**. It explained 74% of the total variability
- The largest coefficients for the second PC were those for FVC and FEV_1/FVC which relate to airflow through the lungs and so it was concluded that this component mainly represented the **degree of airflow obstruction**. It explained a further 15% of the total variability.
- The third PC was dominated by TLCO (transfer factor of the lung for carbon monoxide) and so this component mainly represented **impairment of gas transfer** and explained a further 7% of the total variability
- The fourth PC explained so little of the variability that it was not considered further

- Hence, principal components analysis was able to reduce six lung function variables to three variables (components), where each represented an important, and different, aspect of respiratory morbidity
- The authors used the components in regression analyses to identify men with different forms of lung function abnormalities (see paper for details[1]). In this way just three variables could be used to encapsulate the key features of lung function just as well as the original six variables
- The authors concluded that the principal components method had provided a '**sensitive method of identifying men with unusual lung function**'

Advantages and disadvantages of PC analysis

- A set of inter-correlated variables can be replaced by a smaller set of independent components which represent all of the key features of the original data
- The problems of colinearity in a complex set of predictor variables may be overcome and the role of possible predictor variables can be more easily examined
- Each component is a new variable that is a linear combination of the original variables, and so the actual values of the components are hard to interpret

Reference

1 Cowie H, Lloyd MH, Soutar CA. Study of lung function data by principal components analysis. *Thorax* 1985; **40**(6):438–43.

Cluster analysis

What is cluster analysis?

Cluster analysis is used to identify groups or clusters of individuals who have common features, in terms of known variables. It is has been used to identify groups at high risk of particular adverse events, as a basis for further analysis of causes and prevention. Clustering may be on a single level or may have a hierarchical structure, where groups are identified within groups.

How does cluster analysis work?

The method is used to identify sets of individuals who are more like each other, than they are like other individuals. Since most datasets include several variables on each subject, it is not straightforward to do this with several variables at a time and so there are several methods that can be used. In general, the approaches are based on the following:

- Determining clusters on the basis of measures of how far apart individuals are for quantitative variables
- Determining clusters on the basis of measures of how similar pairs of individuals are

Further details of cluster analysis are beyond the scope of this book but a simple example is given below and references for further reading are listed.

Example

In a study of factors related to premature delivery, researchers used a simple form of cluster analysis on variables associated with early delivery to try to identify groups of women who delivered too early to inform preventive programmes.[1]

The study reported three clusters of women delivering preterm:

- Younger women, predominantly in manual occupations with low income and minimum years of education and with mean gestational age 34.4 weeks
- Older women who smoked, had manual occupations, mainly had low income and minimum years of education and with mean gestational age 33.9 weeks
- Older women who did not smoke, had higher income, more years of education and were less likely to have manual occupations. These women had mean gestational age 35.0 weeks.

The authors concluded that there were 'three subgroups of women delivering preterm: two clusters were predominantly of low social status and the third cluster comprised older women with higher social status who did not smoke'.[1]

Further reading

The fourth edition of Everitt and Landua's *Cluster analysis*[2] has a comprehensive account of cluster methods.

1 Peacock JL, Bland JM, Anderson HR. Preterm delivery: effects of socioeconomic factors, psychological stress, smoking, alcohol, and caffeine. *BMJ* 1995; **311**(7004):531–5.

2 Everitt B, Landua SLM. *Cluster analysis*. 4th ed. London: Edwin Arnold, 2009.

Factor analysis

What is factor analysis?

Factor analysis is related to principal components analysis in that it attempts to reduce the number of variables in a set of data. It is used commonly in the analysis of psychological tests or the analysis of psychological data where the aim is to identify underlying factors.

How factor analysis works

- The underlying hypothesis is that there are a number of common factors that are hidden among the observed data and the method is used to uncover them
- Each observed variable is assumed to be a linear combination of the (unknown) factors
- There is no unique solution to the factor analysis and so a process called **rotation** is used to rotate to a simple structure that is easy to interpret
- Having discovered factors within a set of data, this may need confirming in a further dataset
- As with principal components analysis, a computer program is used for factor analysis

Example

Establishing new dimensions

An example of the use of factors analysis is the well-known Eysenck personality questionnaire (EPQ), which used factor analysis to demonstrate that personality had three dimensions:[1]

- extroversion/introversion
- neuroticism/stability
- psychoticism/socialization

Further reading on factor analysis

- Article on the use of factor analysis in mental health[2]
- Short textbook account of factor analysis (Everitt[3])
- Longer textbook account of factor analysis (Everitt and Dunn[4])

Further reading on multivariate methods

- *Applied multivariate data analysis*[5] gives a thorough account of all methods outlined in this chapter

References

1 Eysenck HJ, Eysenck SBG. *Manual of the Eysenck Personality Inventory*. London: University of London Press, 1964.

2 Ismail K. Unravelling factor analysis. *Evid Based Ment Health* 2008; **11**(4):99–102.

3 Everitt B. *Statistical methods for medical investigations*. 2nd ed. New York: Oxford University Press, 1994.

4 Everitt B, Dunn G. *Applied multivariate data analysis*. London: Edwin Arnold, 1991.