INFORMATION POINT:

Principal component analysis

Principal component analysis (PCA) is amongst the oldest of the multivariate statistical methods of data reduction. It is a method for producing a small number of constructed variables, derived from the larger number of variables originally collected. The idea is to produce a small number of derived variables that are uncorrelated and that account for most of the variation in the original data set. The main reason that we might want to reduce the number of variables in this way is that it helps us to understand the underlying structure of the data.

The derived variables are combinations of the original variables. For example, it might be that students take 10 examinations and some students do well in one exam whilst other students do better in another. It is difficult to compare one student with another when we have 10 marks to consider. One obvious way of comparing students is to calculate the mean score. This is a constructed combination of the existing variables. However, we might get a more useful comparison of overall performances by considering other constructed combinations of the 10 exam marks. PCA is one way of constructing such combinations, doing so in such a way as to account for as much as possible of the variation in the original data. We can then compare students' performance by considering this much smaller number of variables and it might be easier, for example, to identify students that cluster together.

In the preceding paper, the meaning of illness questionnaire consisted of 33 items (variables) and the PCA reduced the dimensions of the data to four constructed variables that are uncorrelated combinations of the 33 original variables. Sometimes the constructed variables are interpretable, as is the case for three of the four variables in the preceding paper, but this will not always be the case. It can often be difficult and dangerous to read too much into the interpretation of the components.

Principal component analysis does not have an underlying statistical model. It is just a mathematical technique and, as such, is used in other statistical analyses that are driven by models, for example, factor analysis. The emphasis in factor analysis is to identify underlying factors that might explain the variability in a large and complex data set. Factor analysis is a two-stage process and PCA is the most commonly used method for the first stage, the extraction of an initial solution. Thus, the mathematical technique of PCA underlies other multivariate statistical methods.

NICOLA CRICHTON

Further reading

Everitt B.S. & Dunn G. (1991) *Applied multivariate data analysis*. Edward Arnold, London.