Principal Component Analysis

Ian Jolliffe

Volume 3, pp. 1580–1584

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Principal Component Analysis

Large datasets often include measurements on many variables. It may be possible to reduce the number of variables considerably while still retaining much of the information in the original dataset. A number of dimension-reducing techniques exist for doing this, and principal component analysis is probably the most widely used of these. Suppose we have n measurements on a vector $\mathbf{x}$ of p random variables, and we wish to reduce the dimension from p to q. Principal component analysis does this by finding linear combinations, $\mathbf{a}_1'\mathbf{x}, \mathbf{a}_2'\mathbf{x}, \ldots, \mathbf{a}_q'\mathbf{x}$, called *principal components*, that successively have maximum variance for the data, subject to being uncorrelated with previous $\mathbf{a}_k'\mathbf{x}$s. Solving this maximization problem, we find that the vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_q$ are the **eigenvectors** of the **covariance matrix**, **S**, of the data, corresponding to the q largest **eigenvalues**. These eigenvalues give the variances of their respective principal components, and the ratio of the sum of the first q eigenvalues to the sum of the variances of all p original variables represents the proportion of the total variance in the original dataset, accounted for by the first q principal components.

This apparently simple idea actually has a number of subtleties, and a surprisingly large number of uses. It was first presented in its algebraic form by **Hotelling** [7] in 1933, though **Pearson** [14] had given a geometric derivation of the same technique in 1901. Following the advent of electronic computers, it became feasible to use the technique on large datasets, and the number and varieties of applications expanded rapidly. Currently, more than 1000 articles are published each year with principal component analysis, or the slightly less popular terminology, principal components analysis, in keywords or title. Henceforth, in this article, we use the abbreviation PCA, which covers both forms. As well as numerous articles, there are two comprehensive general textbooks [9, 11] on PCA, and even whole books on subsets of the topic [3, 4].

An Example

As an illustration, we use an example that has been widely reported in the literature, and which is originally due to Yule et al. [17]. The data consist of scores between 0 and 20 for 150 children aged $4\frac{1}{2}$ to 6 years from the Isle of Wight, on 10 subtests of the Wechsler Pre-School and Primary Scale of Intelligence. Five of the tests were 'verbal' tests and five were 'performance' tests. Table 1 gives the vectors $\mathbf{a}_1$, $\mathbf{a}_2$ that define the first two principal components for these data.

Table 1 Vectors of coefficients for the first two principal components for data from [17]

Variable	$\mathbf{a}_1$	$\mathbf{a}_2$
$\mathbf{x}_1$	0.34	0.39
$\mathbf{x}_2$	0.34	0.37
$\mathbf{x}_3$	0.35	0.10
$\mathbf{x}_4$	0.30	0.24
$\mathbf{x}_5$	0.34	0.32
$\mathbf{x}_6$	0.27	−0.24
$\mathbf{x}_7$	0.32	−0.27
$\mathbf{x}_8$	0.30	−0.51
$\mathbf{x}_9$	0.23	−0.22
$\mathbf{x}_{10}$	0.36	−0.33

The first component is a linear combination of the 10 scores with roughly equal weight (max. 0.36, min. 0.23) given to each score. It can be interpreted as a measure of the overall ability of a child to do well on the full battery of 10 tests, and represents the major (linear) source of variability in the data. On its own, it accounts for 48% of the original variability. The second component contrasts the first five scores on the verbal tests with the five scores on the performance tests. It accounts for a further 11% of the total variability. The form of this second component tells us that once we have accounted for overall ability, the next most important (linear) source of variability in the test scores is between those children who do well on the verbal tests *relative to* the performance tests, and those children whose test score profile has the opposite pattern.

Covariance or Correlation

In our introduction, we talked about maximizing variance and eigenvalues/eigenvectors of a covariance matrix. Often, a slightly different approach is adopted in order to avoid two problems. If our p variables are measured in a mixture of units, then it is difficult to interpret the principal components. What do we mean by a linear combination of weight, height, and temperature, for example? Furthermore, if we measure temperature and weight in degrees Fahrenheit and pounds respectively, we get completely *different principal components* from those obtained from the *same data* but using degrees Celsius and kilograms. To avoid this arbitrariness, we standardize each variable to have zero mean and unit variance. Finding linear combinations of these standardized variables that successively maximize variance, subject to being uncorrelated with previous linear combinations, leads to principal components defined by the eigenvalues and eigenvectors of the **correlation matrix**, rather than the covariance matrix of the original variables. When all variables are measured in the same units, covariance-based PCA may be appropriate, but even here, there can be circumstances in which such analyses are uninformative. This occurs when a few variables have much larger variances than the remainder. In such cases, the first few components are dominated by the high-variance variables and tell us nothing that could not have been deduced by inspection of the original variances. There are certainly circumstances where covariance-based PCA is of interest, but they are not common. Most PCAs encountered in practice are correlation-based. Our example is a case where either approach would be appropriate. The results given above are based on the correlation matrix, but because the variances of all 10 tests are similar, results from a covariance-based analysis would be little different.

How Many Components?

We have talked about q principal components accounting for most of the variation in the p variables. What do we mean by 'most', and, more generally, how do we decide how many components to keep? There is a large literature on this topic – see, for example, [11, Chapter 6]. Perhaps, the simplest procedure is to set a threshold, say 80%, and stop when the first q components account for a percentage of total variation greater than this threshold. In our example, the first two components accounted for only 59% of the variation. We would usually want more than this – 70 to 90% are the usual sort of values, but it depends on the context of the dataset, and can be higher or lower. Other techniques are based on the values of the eigenvalues (for example *Kaiser's rule* [13]) or on the differences between consecutive eigenvalues (the *scree graph* [1]). Some of these simple ideas as well as more sophisticated ones [11, Chapter 6] have been borrowed from **factor analysis**. This is unfortunate because the different objectives of PCA and factor analysis (see below for more on this) mean that, typically, fewer dimensions should be retained in factor analysis than in PCA, so the factor analysis rules are often inappropriate. It should also be noted that, although it is usual to discard low-variance principal components, they can sometimes be useful in their own right, for example in finding outliers [11, Chapter 10] and in quality control [9].

Normalization Constraints

Given a principal component $\mathbf{a}_k'\mathbf{x}$, we can multiply it by any constant and not change its interpretation. To solve the maximization problem that leads to principal components, we need to impose a normalization constraint, $\mathbf{a}_k'\mathbf{a}_k = 1$. Having found the components, we are free to renormalize by multiplying $\mathbf{a}_k$ by some constant. At least two alternative normalizations can be useful. One that is sometimes encountered in PCA output from computer software is $\mathbf{a}_k'\mathbf{a}_k = l_k$, where l_k is the kth eigenvalue (variance of the kth component). With this normalization, the jth element of $\mathbf{a}_k$ is the correlation between the jth variable and the kth component for correlation-based PCA. The normalization $\mathbf{a}_k'\mathbf{a}_k = 1/l_k$ is less common, but can be useful in some circumstances, such as finding **outliers**.

Confusion with Factor Analysis

It was noted above that there is much confusion between principal component analysis and factor analysis. This is partially caused by a number of widely used software packages treating PCA as a special case of factor analysis, which it most certainly is not. There are several technical differences between PCA and factor analysis [10], but the most fundamental difference is that factor analysis explicitly specifies a model relating the observed variables to a smaller set of underlying unobservable factors. Although some authors [2, 16] express PCA in the framework of a model, its main application is as a descriptive, exploratory technique, with no thought of an underlying model. This descriptive nature means that distributional assumptions are unnecessary to apply PCA in its usual form. It can be used, although an element of caution may be needed in interpretation, on discrete, and even binary, data, as well as continuous variables. One notable feature of factor analysis is that it is generally a two-stage procedure; having found an initial solution, it is rotated towards *simple structure* (*see* **Factor Analysis: Exploratory**). The purpose of *factor rotation* (*see* **Factor Analysis: Exploratory**) is to make the coefficients or loadings relating variables to factors as simple as possible, in the sense that they are either close to zero or far from zero, with few intermediate loadings. This idea can be borrowed and used in PCA; having decided to keep q principal components, we may rotate within the q-dimensional subspace defined by the components in a way that makes the axes as easy as possible to interpret. This is one of a number of techniques that attempt to simplify the results of PCA by postprocessing them in some way, or by replacing PCA with a modified technique [11, Chapter 11].

Uses of Principal Component Analysis

The basic use of PCA is as a dimension-reducing technique whose results are used in a descriptive/exploratory manner, but there are many variations on this central theme. Because the 'best' two- (or three-) dimensional representation of a dataset in a least squares sense (*see* **Least Squares Estimation**) is given by a plot of the first two- (or three-) principal components, the components provide a 'best' low-dimensional graphical display of the data. A plot of the components can be augmented by plotting variables as well as observations on the same diagram, giving a **biplot** [6].

PCA is often used as the first step, reducing dimensionality before undertaking another multivariate technique such as cluster analysis (*see* **Cluster Analysis: Overview**) or **discriminant analysis**. Principal components can also be used in **multiple linear regression** in place of the original variables in order to alleviate problems with *multicollinearity* [11, Chapter 8]. Several dimension-reducing techniques, such as **projection pursuit** [12] and **independent component analysis** [8], which may be viewed as alternatives to PCA, nevertheless, suggest preprocessing the data using PCA in order to reduce dimensionality, before proceeding to the technique of interest. As already noted, there are also occasions when low-variance components may be of interest.

Extensions to Principal Component Analysis

PCA has been extended in many ways. For example, one restriction of the technique is that it is linear. A number of nonlinear versions have, therefore, been suggested. These include the 'Gifi' [5] approach to **multivariate analysis**, and various nonlinear extensions that are implemented using **neural networks** [3]. Another area in which many variations have been proposed is when the data are **time series**, so that there is dependence between observations as well as between variables [11, Chapter 12]. A special case of this occurs when the data are functions, leading to *functional data analysis* [15]. This brief review of extensions is by no means exhaustive, and the list continues to grow – see [9, 11].

References

[1] Cattell, R.B. (1966). The scree test for the number of factors, *Multivariate Behavioral Research* **1**, 245–276.

[2] Caussinus, H. (1986). Models and uses of principal component analysis: a comparison emphasizing graphical displays and metric choices, in *Multidimensional Data Analysis*, J. de Leeuw, W. Heiser, J. Meulman and F. Critchley eds, DSWO Press, Leiden, pp. 149–178.

[3] Diamantaras, K.I. & Kung, S.Y. (1996). *Principal Component Neural Networks Theory and Applications*, Wiley, New York.

[4] Flury, B. (1988). *Common Principal Components and Related Models*, Wiley, New York.

[5] Gifi, A. (1990). *Nonlinear Multivariate Analysis*, Wiley, Chichester.

[6] Gower, J.C. & Hand, D.J. (1996). *Biplots*, Chapman & Hall, London.

[7] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology* **24**, 417–441, 498–520.

[8] Hyvärinen, A., Karhunen, J. & Oja, E. (2001). *Independent Component Analysis*, Wiley, New York.

[9] Jackson, J.E. (1991). *A User's Guide to Principal Components*, Wiley, New York.

[10] Jolliffe, I.T. (1998). Factor analysis, overview, in *Encyclopedia of Biostatistics* Vol. 2, P. Armitage & T. Colton, eds, Wiley, New York 1474–1482.

[11] Jolliffe, I.T. (2002). *Principal Component Analysis*, 2nd Edition, Springer, New York.

[12] Jones, M.C. & Sibson, R. (1987). What is projection pursuit? *Journal of the Royal Statistical Society, A* **150**, 1–38. (including discussion).

[13] Kaiser, H.F. (1960). The application of electronic computers to factor analysis, *Educational and Psychological Measurement* **20**,141–151.

[14] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space, *Philosophical Magazine* **2**, 559–572.

[15] Ramsay, J.O. & Silverman, B.W. (2002). *Applied Functional Data Analysis, Methods and Case Studies*, Springer, New York.

[16] Tipping, M.E. & Bishop, C.M. (1999). Probabilistic principal component analysis, *Journal of the Royal Statistical Society, B* **61**, 611–622.

[17] Yule, W., Berger, M., Butler, S., Newham, V. & Tizard, J. (1969). The WPPSI: an empirical evaluation with a British sample, *British Journal of Educational Psychology* **39**, 1–13.

(*See also* **Multidimensional Scaling**)

IAN JOLLIFFE