J Neurol (1990) 237:461–464

Journal of
Neurology
© Springer-Verlag 1990

Multivariate data reduction by principal components, with application to neurological scoring instruments*

J. A. Koziol[1] **and W. Hacke**[2]

[1]Department of Molecular and Experimental Medicine, Research Institute of Scripps Clinic, 10666 North Torrey Pines Road, La Jolla, CA 92037, USA
[2]Neurologische Klinik, Universität Heidelberg, Im Neuenheimer Feld 400, W-6900 Heidelberg 1, Federal Republic of Germany

Received November 27, 1989 / Received in revised form July 16, 1990 / Accepted August 2, 1990

Summary. Principal components analysis is widely used as a practical tool for the analysis of multivariate data. The aim of this analysis is to reduce the dimensionality of a multivariate data set to the smallest number of meaningful and independent dimensions. The analysis can also provide interpretable linear functions of the original measured variables that may serve as valuable indices of variation. A brief introduction to principal components analysis is given herein, followed by an examination of a particular set of multivariate data accruing from a study of acute brain injuries in a pediatric population, in which severity of brain injury had been assessed with the Glasgow Coma Scale (CGS). Principal components analysis reveals that the GCS sum score is a particularly inefficient summarizer of information in this cohort. The determination of an objective weighting of measured variables, as provided through principal components analysis, is essential in the construction of meaningful neurological scoring instruments.

Key words: Neurological scoring instruments – Glasgow Coma Scale – Principal components analysis – Coma

Introduction

The analysis of multivariate data can be a difficult and frustrating problem. One frequently used approach is to reduce the dimensionality of the data, so as to engender easier understanding, visualization, and interpretation. This reduction in dimensionality can entail some loss of information that may be in the data; hence, any reduction ought to retain sufficient detail for adequate representation of the original data. The purpose of this note is to describe principal components analysis, a classical statistical reduction technique introduced by Karl Pearson in 1901 [4] and further developed by Hotelling [1]. We first address the issues of dimensionality reduction and information loss with principal components and then consider neurological scoring instruments in this context.

Principal components analysis

We begin with the familiar representation of the density function

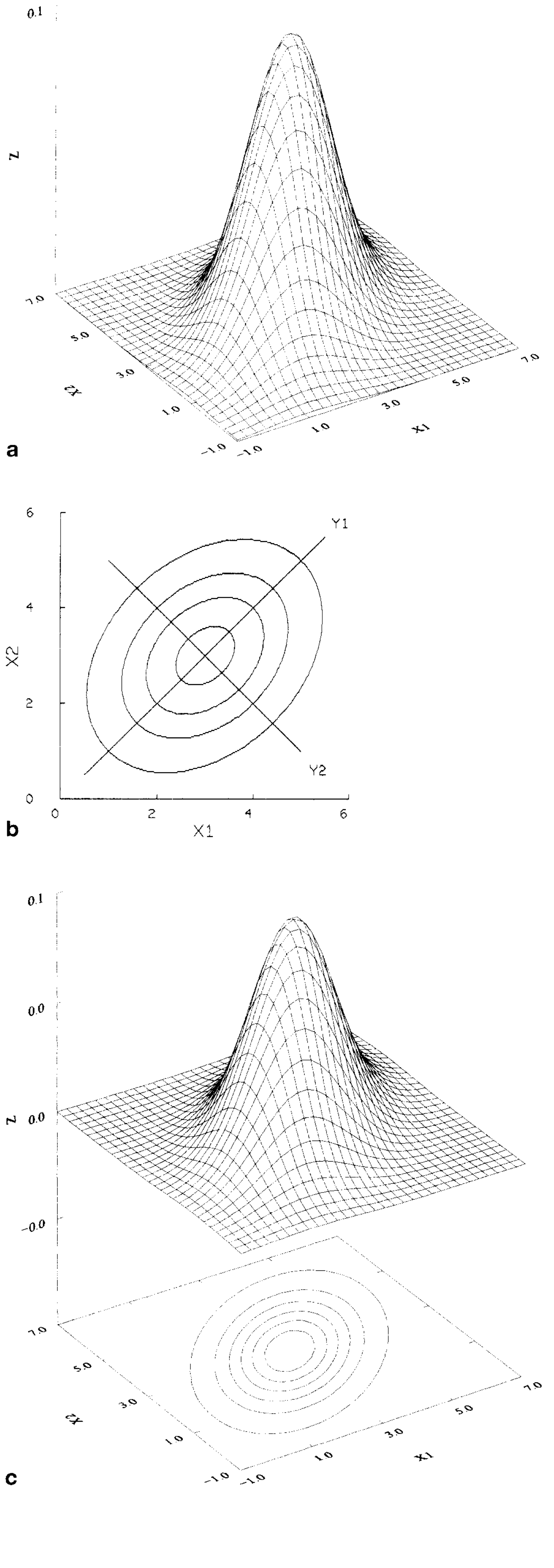
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/2\sigma^2}$$

of the univariate normal distribution. The density function is symmetrical about the value μ, a location parameter which represents the mean of the distribution; a second parameter, the variance, designated σ^2, represents a measure of spread or dispersion. For example, about 68.3% of the density mass lies within 1 SD $\sigma = \sqrt{\sigma^2}$ of the mean μ, and about 95.4% lies within 2 SD of μ. The normal distribution is an idealization, but serves as an adequate approximation for many random variables observed in practice: indeed, by the central limit theorem in probability theory, although a random sample may come from a non-normal population, the sample mean $\bar{x}$ will typically be approximately normally distributed for large sample sizes.

Perhaps less familiar is the notion of a bivariate (and more generally, multivariate) normal distribution. A multivariate probability distribution can be defined completely by specifying the distributions of all linear combinations of its components. Thus, the random variables X_1 and X_2 have jointly a bivariate normal distribution if and only if all linear combinations $aX_1 + bX_2$ are univariate normally distributed, for any prespecified constants a and b. In Fig. 1a we give the density function of a bivariate normal distribution for X_1 and X_2, when X_1 is

*This work was funded in part by the Deutsche Forschungsgemeinschaft (grant Ha 1394/2-1), National Institutes of Health (grants RR00833 and CA41582), and an award from the Alexander von Humboldt Foundation

Offprint requests to: W. Hacke



univariate normal with mean $\mu_1 = 3$ and variance $\sigma_1^2 = 2$, X_2 is univariate normal with mean $\mu_2 = 3$ and variance $\sigma_2^2 = 2$, and the covariance of X_1 and X_2, denoted σ_{12}, is 1. Note that an additional parameter, the covariance, is needed to characterize the bivariate normal distribution; a related term, the correlation of X_1 and X_2, denoted by ρ, is dimensionless, and equals $\sigma_{12}/\sqrt{\sigma_1^2 \sigma_2^2}$. We may easily show that $-1 \le \rho \le 1$; if $\rho = 0$, then X_1 and X_2 are independent, and if $\rho = \pm 1$, then X_1 and X_2 are perfectly linearly correlated.

It is convenient when working in higher dimensions to summarize the parameters in vector and matrix notation. In p dimensions, the random vector X consists of p component univariate random variables, denoted X_1, $X_2, \ldots, X_p$; the mean μ will be a $p \times 1$ vector, with the component μ_i being the mean of X_i, $i = 1, 2, \ldots, p$; and the covariance matrix Σ will be a $p \times p$ matrix, with elements σ_{ij}, $i, j = 1, 2, \ldots, p$. The ith diagonal element of Σ, σ_{ii}, is the variance of the ith component of X, X_i; we sometimes denote this by σ_i^2, as in the bivariate case. The correlation coefficient between X_i and X_j is defined as

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\,\sigma_{jj}}} = \frac{\sigma_{ij}}{\sigma_i\,\sigma_j}.$$

Consider next a plane in Fig. 1a parallel to the $X_1 - X_2$ plane that slices through the density function. This intersection would form an ellipse, as depicted in Fig. 1b: that is, the normal distribution in two dimensions has constant density on ellipses of the form $(X - \mu)^T\, \Sigma^{-1}\, (X - \mu) = c^2$, c being a constant. [Here, $(X - \mu)^T$ denotes the transpose of the vector $(X - \mu)$, and Σ^{-1} denotes the inverse of the covariance matrix Σ. $(X - \mu)^T\, \Sigma^{-1}\, (X - \mu) = c^2$ is simply a linear algebraic rendition of the equation for an ellipse.] These ellipses (or ellipsoids in higher dimensions) are called the contours of the distribution, or the ellipses of equal concentration. If the mean vector μ is the null vector (all zeros), these contours are centered at the origin; and, when Σ is diagonal (in the bivariate case, if X_1 and X_2 are independent, so that $\sigma_{12} = 0$), the contours are circles (or in higher dimensions spheres or hyperspheres). Figure 1c relates the density and the contours of the bivariate normal distribution considered previously, from a different perspective.

Suppose now that the $p \times 1$ vector $X = (X_1, X_2, \ldots, X_p)^T$ has an arbitrary multivariate distribution. The principal components of this distribution are uncorrelated linear combinations of the original variates $X_1, X_2, \ldots, X_p$ which successively account for the most part of variation in a population. To make this definition meaningful, we must indicate how to assess variation in the multivariate setting. We have already noted that the matrix Σ consists of parameters that are related to the univariate notion of variance. However, there also exist two com-

Fig. 1. **a** Density function of the bivariate normal distribution with mean vector $\mu = \binom{3}{3}$, and covariance matrix $\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$. **b** Ellipses of equal concentration of the bivariate normal distribution pictured in **a**, showing the directions Y1 and Y2 of the principal components. **c** Composite of **a** and **b**, with ellipses of equal concentration projected below the bivariate density function monly used univariate quantities that measure multivariate scatter:

1. The generalized variance of Σ, the determinant of Σ, denoted $|\Sigma|$;

2. The total variation, trace (Σ), the sum of the diagonal elements of Σ.

For both measures, large values indicate a high degree of scatter of the distribution of X about its mean vector μ, and low values indicate tight concentration about μ. Now, the first principal component is the normalized linear combination of the original variates with maximum variance; the second principal component is that normalized linear combination which is uncorrelated with the first principal component and has maximal variance (and so on, if we are working in dimensions higher than 2. We may find p principal components of a non-singular p-dimensional distribution). Algebraically, principal components analysis is equivalent to an orthogonal transformation of the original set of variables into a set of new variables which are uncorrelated with each other, and which are ordered in terms of decreasing variance. Moreover, the sum of the variances of the new variates equals the total variation, trace (Σ), and the product of the variances equals the generalized variance, $|\Sigma|$.

Geometrically, the principal components transformation facilitates interpretation of the ellipsoids of equal concentration for the multivariate normal distribution, because the principal components represent the major and minor semi-axes of the ellipsoids, as noted in Fig. 1b. Principal components are often used to replace the original variates by a smaller number of uncorrelated linear combinations of them without incurring much loss of information. For example, if there is little scatter in the second principal component direction Y_2 relative to the first principal component direction Y_1 in Fig. 1b, little information is lost by simply discarding Y_2, and thereby reducing a two-dimensional distributional problem to one dimension. More generally, the proportion of total variation explained by, say, the first k principal components gives a quantitative measure of the amount of information retained in the reduction from p to k dimensions. A rule of thumb for excluding or discarding principal components is to include just enough components to explain, say, 80–90% of the total variation. This attempt to reduce dimensionality is often described as parsimonious summarization of the data; the method of principal components analysis affords the means of determining whether such summarization has or even can be successfully achieved with particular data sets. We refer the reader to any of several recent books on multivariate statistical methods (e.g. [2]) for more detailed treatment of principal components analysis and related multivariate techniques.

Application to neurological scoring instruments

For concreteness, let us consider the Glasgow Coma Scale (GCS), perhaps the most frequently used method for measuring overall responsiveness in patients with acute cerebral disorders. The GCS is a clinical scale developed by Teasdale and colleagues [5, 6] for assessing depth and duration of impaired consciousness and coma. It comprises independent determinations of three aspects of behavior: eye response (1-2-3-4 scale), motor responsiveness (1-2-3-4-5-6 scale), and verbal performance (1-2-3-4-5 scale), which when summed yield an overall numerical rating between 3 and 15. Initially, the authors had recommended using the information of the three subscores by means of a profile, but in subsequent publications the use of a sum score was emphasized.

Our aim here is to examine whether the GCS sum score adequately summarizes the information available in the three individual scales. For illustrative purposes we focus on the investigation by Kraus et al. [3], who reported on the nature, clinical course, and early outcomes in a cohort of 709 pediatric patients suffering brain injury. Severity of brain injury was measured with the GCS. We are grateful to Dr. Kraus for making these data available to us for principal components analysis.

The numbers of individuals classified into the various categories on each scale follow a multinomial distribution, for which the multivariate normal distribution provides an adequate large sample approximation (via the central limit theorem mentioned above). (Note, however, that many of the properties of principal components are not dependent upon the assumption that the data follow a multivariate normal distribution.) The GCS scoring scheme on each scale is equivalent to taking a linear combination of the outcome events; the information content in this linear combination may be compared with the maximal information content available in the optimal linear combination, the first principal component, in terms of total variation. Similarly, the overall GCS sum score represents another linear combination of the outcome events considered jointly; the information content in this linear combination may be compared with that of the first principal component from the joint distribution of the eye, motor, and verbal scales. We compare the GCS scores and the corresponding first principal components in Table 1, which is derived from individuals in the Kraus study with GCS sum scores less than 14 (so as to eliminate individuals who arguably presented with non-severe brain injuries).

It is apparent from Table 1 that, as might have been anticipated, the GCS scoring scheme preserves a distressingly low proportion of the total variation on each of the scales: the GCS scoring scheme accounts for 8.5% of the total variation on the motor scale, 9.1% on the verbal scale, and 12.9% on the eye scale. Moreover, even if we were to optimize the assignment of scores, by means of the first principal component that maximizes the proportion of total variation explained, we would not be successful: with the first principal component, we are accounting for merely 46.3–61.5% of the total variation on the scales. Even though this constitutes a substantial improvement over the GCS scoring scheme, we are nevertheless incurring a measurable loss of information by such reduction of multivariate data to a univariate scale. This, of course, is not at all surprising, as there is no reason to expect solely one linear combination to capture most of the relevant information present in a multi-

Table 1. Information content of Glasgow Coma Scale (GCS) scores and first principal component (FPC) in a study of pediatric brain injuries

Scale	Proportion of total variation explained by	
	GCS score	FPC
Eye	0.129	0.615
Motor	0.085	0.463
Verbal	0.091	0.552
Overall sum	0.086	0.444

variate setting. This obviously pertains to the sum score also: much information is lost in reducing different clinical profiles to GCS sum scores, as the sum score accounts for a mere 8.6% of the total variation on the three scales simultaneously. The first principal component, by comparison, captures 44.4% of the total variation, but this again represents an intolerably large information loss. Even the first three principal components jointly account for less than 70% of the total variation on the three scales simultaneously: principal components analysis with these data demonstrates the futility of attempting to summarize the multivariate profiles in a univariate manner.

Discussion

The introduction of scales or scores for the assessment of neurological function represents operationally a reduction of multivariate data to a univariate quantity. As mentioned before, the GCS was initially constructed to assess three independent aspects of functions that can be compromised in the presence of decreased consciousness. The use of a sum score was not recommended in the early publications. Furthermore, the presently 3–15 for the sum score had originally been 0–12 in earlier publications, indicating that the score values were subject to change in the initial years after their introduction. A number of additional problems are associated with this scoring instrument:

1. The score values are not distributed equally, thus preventing most statistical approaches from being conventionally performed with the results of the GCS.

2. There exists the real possibility of "pseudoscoring," particularly with the verbal performance scale, if the type of injury or iatrogenic intervention (such as intubation and artificial ventilation) precludes actual testing of verbal performance.

3. Not all GCS values represent states that may be considered "coma": indeed, only two of the four items of the "eye response" subscore and two of the five "verbal response" items refer to conditions present in "coma".

4. Issues of reliability, validity and interobserver variability need to be addressed for a clinical instrument so widely used as the GCS, but are largely missing.

5. Finally, the GCS was initially designed not as a scoring instrument in the psychological sense, but as a data-analytic aid for the retrospective classification of patients.

The fundamental question arises as to whether the reduction of the information to a sum score effects an adequate representation of the information available with the original data. Principal components analysis presents us a solution to the problem of determining an optimal reduction if our data are multivariate normally distributed, namely, choose the first principal component: this represents the major axis of the ellipsoids of concentration of our data, and will have maximal variance of any normalized linear combination of our original variates. Moreover, we may quantify the proportion of the total variation of our original data that is preserved by the first, or any set of, components. A "successful" principal components analysis achieves a reduction in dimensionality from p to k dimensions, while simultaneously preserving about 80–90% of the total variation. Rarely will merely the first principal component suffice for this task – a fact that should caution against the indiscriminate reliance on univariate scores for multivariate data summarization.

Let us also remark that, if one nevertheless wishes to derive a univariate score for quantification, a further word of caution is in order: the first principal component will yield the optimal linear combination (that is, a scoring scheme for categorical data), but this is a data-dependent derivation. This is, the optimal scores will depend on the particular population under investigation, and will not in general be immediately applicable to a different clinical population. Prospective validation is essential before any particular assignment of scores can be universally recommended.

In summary, for a number of reasons we discourage strongly the use of the GCS sum score. In addition to apparent weaknesses in the scale construction, validation and reliability, the amount of information loss incurred by simply summing across the individual scale scores can often be enormous. We instead recommend that the GCS be used in the way it was originally designed, namely, as a three-dimensional profile on eye, motor, and verbal responses.

References

1. Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24:417–441, 498–520
2. Johnson RA, Wichern DW (1989) Applied multivariate statistical analysis, 2nd edn. Prentice-Hall, New Jersey
3. Kraus JF, Fife D, Conroy C (1987) Pediatric brain injuries: the nature, clinical course, and early outcomes in a defined United States population. Pediatrics 79:501–507
4. Pearson K (1901) On lines and planes of closest fit to systems of points in space. Phil Mag 2:559–572
5. Teasdale G, Jennett B (1974) Assessment of coma and impairment of consciousness. Lancet II:81–84
6. Teasdale G, Jennett B (1976) Assessment and prognosis of coma after head injury. Acta Neurochir (Wien) 34:45–55