Principal component analysis and exploratory factor analysis

IT Joliffe and Bjt Morgan

Stat Methods Med Res 1992 1: 69
DOI: 10.1177/096228029200100105

Statistical Methods in Medical Research 1992 **1**: 69–95

Principal component analysis and exploratory factor analysis

IT Joliffe Department of Mathematical Sciences, University of Aberdeen and **BJT Morgan** Institute of Mathematics and Statistics, University of Kent

In this paper we compare and contrast the objectives of principal component analysis and exploratory factor analysis. This is done through consideration of nine examples. Basic theory is presented in appendices. As well as covering the standard material, we also describe a number of recent developments. As an alternative to factor analysis, it is pointed out that in some cases it may be useful to rotate certain principal components if and when that is appropriate.

1 Introduction

In medical research, as in many other fields, large numbers of variables are sometimes collected on each of a number of individuals. In such cases the sheer volume of data can be daunting, and it is a natural objective to try to reduce the number of variables whilst preserving as much of the original information as possible. Principal component analysis and factor analysis provide two ways of tackling this objective.

In this paper we shall illustrate the discussion of these two techniques through a number of real examples, most of which are medical, and many of which we have encountered in the course of our own work. The first example describes a set of individuals treated for an extradural haematoma at the Brook Hospital, Woolwich. It differs from most of our other examples in that it is more extensive than data sets which are typically used to illustrate multivariate analysis techniques. However, it is still much smaller than many sets of medical data. We shall now present this example, and use it to motivate the introductory material on principal component analysis. Further details of the results of principal component and factor analyses applied to this example are deferred until Section 3.

Example 1: Extradural haematoma

The data set of this example was analysed in an unpublished dissertation at the University of Kent.[1] On each of 172 patients with an extradural haematoma, treated at the Brook Hospital, a number of variables were measured. We focus on 12 of these, viz., age, sex, interval between injury and treatment, the result of special tests (e.g. whether or not the skull was fractured), whether or not alcohol was present, whether general anaesthesia was used in the operation, best verbal response on admission to hospital (VADM), best verbal response prior to operation (VP), best motor response on admission to hospital (MADM), best motor response prior to operation (MP), distance, in miles, between referring hospital and the Brook, and outcome following operation. The verbal and motor responses record coma state, and were measured on a scale from 1 (best: orientated and obeying commands), to 5 (worst: no response of either kind) and outcome was recorded on a scale from 1 (full recovery) to 6 (dead).

An extradural haematoma may result from a knock on the head. Early diagnosis is

Address for correspondence: Ian T Joliffe, Department of Mathematical Sciences, University of Aberdeen, The Edward Wright Building, Dunbar Street, Aberdeen A89 2TY, UK.

difficult, since the injured individual may appear quite normal under observation in a local (referring) hospital, but then rapidly deteriorates. Typically, as complications occur, the patient is transferred to the Regional Neurological Unit (in this case the Brook Hospital) for surgery. Because of interest in this particular injury at the Brook Hospital, the records are complete and there are no missing data. More detail is given in unpublished work by Bartlett *et al.*[2] and Davenport-Ellerby.[1] As can be appreciated from the description above, the variables are of several different kinds. This data set has been used in a range of analyses, with for instance discriminant analysis being applied to the outcome category, which clearly has a different status from the other variables. When we analyse these data in Section 3, this distinction is made quite clear. The correlation matrix for the variables is given in Table 1. Clearly there are some high correlations present, notably between certain of the coma state variables. Additionally, positive and negative correlations are present.

Table 1 Correlation matrix for the variables measured in the extra-dural haematoma study

	Age	Sex	Interval	Tests	Alcohol	Anaesthetic	VADM	VP	MADM	MP	Distance
Age	1										
Sex	0.004	1									
Interval	−0.020	0.034	1								
Tests	0.039	0.008	0.236	1							
Alcohol	0.153	−0.101	0.086	0.116	1						
Anaesthetic	−0.181	0.088	0.180	0.108	0.040	1					
VADM	0.179	−0.065	−0.141	−0.040	0.305	−0.028	1				
VP	0.062	0.013	−0.598	−0.296	−0.037	−0.162	0.183	1			
MADM	0.175	−0.023	−0.197	−0.126	0.236	−0.088	0.838	0.211	1		
MP	0.070	−0.021	−0.561	−0.314	−0.040	−0.227	0.224	0.760	0.299	1	
Distance	0.077	0.061	0.165	0.114	−0.109	0.095	−0.047	−0.065	−0.090	−0.065	1
Outcome	0.390	0.002	−0.251	−0.131	0.099	−0.292	0.285	0.390	0.349	0.549	0.023

Standard notation for multivariate data tables uses p as the number of variables measured per patient (here $p = 12$) and n for the number of patients (here $n = 172$), with rows of the table corresponding to patients, and columns to variables.

There are two ways in which we might attempt to reduce the number of variables and hence the dimensionality, p, of a data set. First, we could simply choose a subset of m of the original variables. For example, in the brain data set we might construct a range of two-way tables, presenting the relationships between age and outcome, coma-state variables and outcome, and so on. Because, e.g. MADM and VADM are highly correlated, we may feel that we need only consider one of them in our tables. Similarly, for other pairs of correlated variables. However three-way and higher-order tables will still need to be considered, and variable selection in this way is rather arbitrary and subjective.

An alternative strategy is to construct m *new* or derived variables from the p original ones. Each of the new m variables is now, in general, different from any of the original variables. This approach is perhaps less appealing, intuitively, than retaining a subset of the original variables. However, it has the advantage that for the same dimensionality, m, we are usually able to retain a greater proportion of the original variation by using derived variables than we can with a simple subset of variables. Both principal component analysis (PCA) and factor analysis (FA) are dimension-reducing techniques which use the idea that a small number of derived or underlying variables can replace the originally measured variables with little loss of information. Despite their similarity of purpose, there are fundamental differences in the reasoning behind the two techniques. Recent examples of FA and PCA in the medical literature show that there is confusion, not only over the differences between the two methods, but also regarding exactly what each technique actually does. A major objective of this article is to clarify these issues and show how PCA and FA can be used in medical applications.

Section 2 of the paper explains the reasoning behind each technique in some detail, although some technical material is relegated to an Appendix. This section highlights differences, but also similarities in implementation, between the two methods. It also gives a brief history of the techniques, together with (again briefly) some recent examples from the medical literature.

Section 3 discusses, in detail, the application of both techniques to the brain data set of Example 1. Both Sections 2 and 3 will be restricted to the most straightforward forms of PCA and FA. It is interesting, however, to see just how many questions of detail need to be answered in an actual implementation of either technique. A number of these questions are addressed in Section 4, and some of the points made in this section again serve to highlight the differences between FA and PCA. Section 5 continues the theme of Section 4, but discusses and illustrates some modifications of, and extensions to, PCA and FA. General conclusions are drawn in Section 6.

2 The techniques defined

2.1 Principal component analysis

As noted above, PCA aims to replace p measured variables $x_1, x_2, \ldots, x_p$, by m derived variables, $z_1, z_2, \ldots, z_m$, where m is much less than p, whilst minimizing any loss of information. The variables $z_1, z_2, \ldots$ are taken to be *linear* functions of $x_1, x_2, \ldots, x_p$ with the following properties:

a) z_1 has maximum possible variance among all possible linear functions

$$z = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_p x_p$$

of $x_1, x_{27} \ldots, x_p$.

b) z_2 has maximum possible variance among all possible linear functions of $x_1, x_2, \ldots, x_p$, subject to z_2 being uncorrelated with z_1.

c) In general, z_k has maximum possible variance among all possible linear functions of $x_1, x_2, \ldots, x_p$, subject to z_k being uncorrelated with $z_1, z_2, \ldots, z_{k-1}$, for $2 \leqq k \leqq p$.

The derived variables $z_1, z_2, \ldots, z_p$ are the *principal components* (PCs). While there are, in fact, as many PCs (p) as there are variables, typically we can account for much of the variation in the original *x*s with far fewer PCs. The definition above guarantees that the PCs successively account for as much of the total variation in the data set, measured by variance, as possible. At the same time, the requirement that PCs are uncorrelated ensures that each component accounts for a distinguishable source of variation.

PCA can be defined in a number of alternative, equivalent ways, but the definition above is the most usual. There are also a number of points of detail on which the definition may vary. We leave details of most of these variations until Section 4, but two need to be mentioned here. First, the definition above does not have a solution without imposing some restriction. This is because we could, for example, for any potential solution multiply each of $\alpha_1, \alpha_2, \ldots, \alpha_p$ by 10, which increases the variance by a factor of 100. Thus the variance of linear functions is unbounded, unless we impose some normalization constraint on the αs. If the kth PC is now written as $z_k = \alpha_{k1}x_1 + \ldots + \alpha_{kp}x_p = \boldsymbol{\alpha}_k'\mathbf{x}$, using standard vector notation, then the constraint which is used in the derivation of the Appendix A.1 is that

$$\sum_{j=1}^{k} \alpha_{kj}^2 = 1,$$

or equivalently, $\boldsymbol{\alpha}_k'\boldsymbol{\alpha}_k = 1$. We shall discuss alternative normalization constraints in Section 4.3

The second detail that we need to mention is to make the assumption that the original variables have been standardized to have unit variance before we attempt to find PCs. It is irrelevant whether or not the variables are standardized to have zero mean. However, it is conventional also to standardize variables to have zero means. Resulting components then also have zero means, as can be seen from the figures of this paper. In Appendix A.1 it is noted that we calculate PCs by finding eigenvalues and eigenvectors of a matrix. If we use standardized variables, then the matrix concerned is the correlation matrix of the original variables, whereas for unstandardized variables, we consider the corresponding covariance matrix instead. Whether it is more appropriate to use the covariance matrix or correlation matrix depends on the data set being analysed. However, it is far more common for the correlation matrix to be preferred. This is because analysis of the covariance matrix will only make sense if all variables are measured in the same units, with a similar range of variation for each variable. In the majority of studies, different variables will be in different units, and this is clearly the case for the data of Example 1. When a principal component analysis is carried out for the covariance matrix for the brain data, the first two components are, respectively, a weighted average of the age and distance measures, and a contrast of these, describing 53% and 45% of the total variance respectively. Such a result is not surprising since the age and distance variables have by far the largest variances out of the set of 12 variables. In this case the principal component analysis is of questionable use.

Even when all the variables are measured in the same units, if the variances of the variables differ considerably then the PCs are entirely predictable, and provide no useful information beyond that of the relative sizes of the variances. Jolliffe[3] gives an example where the data consist of eight blood chemistry variables for 72 patients, which illustrates this point. Section 2.3 of the same book discusses the choice between covariance and correlation matrices more fully, but with one exception we shall assume from now on that a correlation matrix is being used. The exception arises in connection with our consideration of metric scaling in Section 5.

The discussion so far has concentrated on the p original (standardized) variables $x_1, x_2, \ldots, x_p$. In fact, a typical data set will consist of n observations on each variable, often measurements of p variables on n patients, as in the brain data set. The results of a PCA on the data matrix with elements x_{ij}, the jth measurement on the ith patient, have three aspects to them.

i) The variances $\lambda_1, \lambda_2, \ldots, \lambda_p$ of the principal components. λ_k is the kth largest eigenvalue of the correlation matrix, $\mathbf{R}$, for our data set – see Appendix A.1.

ii) The coefficients or loadings, α_{kj}, of the jth variable in the kth PC; the vector $\boldsymbol{\alpha}_k$

consisting of elements α_{kj} is the eigenvector corresponding to λ_k.

iii) The values, z_{ik} of each component for each patient (the PC scores), where

$$z_{ik} = \sum_{j=1}^{p} \alpha_{kj} x_{ij}.$$

Each of these three aspects is useful in analysing the original data set. The variances, λ_k, are crucial in determining how great a reduction in dimensionality can be achieved, whilst retaining much of the original variation. The ratio

$$\sum_{k=1}^{m} \lambda_k \Big/ \sum_{k=1}^{p} \lambda_k$$

represents the proportion of the total variation accounted for by the first m PCs. If this is large for relatively small m, it means that a substantial reduction in dimensionality is possible. We have already seen that for the brain data just two dimensions sufficed to describe the variance structure for the unscaled variables. However, when we consider the case of the scaled variables, as many as six components still only describe 77% of the total variance.

The coefficients α_{kj} allow us to interpret the sucessive components in terms of the original variables. The hope is that the pattern of coefficients in the first few PCs will be sufficiently simple to allow easy interpretation of what these components represent. As we shall see in Section 3, interpretation of principal component coefficients is often not straightforward. It should also always be borne in mind that the coefficients are themselves realizations of random variables. This point is graphically made in a bootstrap study of the possible variation in component coefficients, by Diaconis and Efron.[4] Example 2 which follows is unusual in having a nicely-defined structure present for all of the components.

Example 2: Strengths of reflexes in patients with diabetes

The data of this example were provided by Pfizer Central Research, Sandwich, and are discussed by Jolliffe[3] (pp. 47–49). The strengths of reflexes at 10 different sites of the body were measured for each of 143 individuals with diabetes. The correlation matrix is given in Table 2.

Table 2 Correlation matrix for ten variables measuring reflexes. See text for the location of each site

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1	1.00									
S2	0.98	1.00								
S3	0.60	0.62	1.00							
S4	0.71	0.73	0.88	1.00						
S5	0.55	0.57	0.61	0.68	1.00					
S6	0.55	0.57	0.56	0.68	0.97	1.00				
S7	0.38	0.40	0.48	0.53	0.33	0.33	1.00			
S8	0.25	0.28	0.42	0.47	0.27	0.27	0.90	1.00		
S9	0.22	0.21	0.19	0.23	0.16	0.19	0.40	0.41	1.00	
S10	0.20	0.19	0.18	0.21	0.13	0.16	0.39	0.40	0.94	1.00

As in Table 1, we see a number of large correlations, but in contrast to Table 1, there are no negative correlations. When all correlations are positive we might expect the first principal component to be a weighted average of the 10 variables, in this case separating people out according to average reflex strength. In such cases the other principal components are often of greater interest. This situation is quite a common one, occurring when variables measure size, or examination performance, incidence of facial spots of different kinds (Example 9 discussed later), etc. Here the 10 sites are:

S1	Right tricep
S2	Left tricep
S3	Right bicep
S4	Left bicep
S5	Right wrist
S6	Left wrist
S7	Right knee
S8	Left knee
S9	Right ankle
S10	Left ankle

The principal component coefficients are given in Table 3.

Table 3 Principal components based on the correlation matrix of Table 2

	Component number									
	1	2	3	4	5	6	7	8	9	10
	coefficients									
S1	0.3	−0.2	0.2	−0.5	0.3	0.1	−0.1	−0.0	−0.6	0.2
S2	0.4	−0.2	0.2	−0.5	0.3	0.0	−0.1	−0.0	0.7	−0.3
S3	0.4	−0.1	−0.1	−0.0	−0.7	0.5	−0.2	0.0	0.1	0.1
S4	0.4	−0.1	−0.1	−0.0	−0.4	−0.7	0.3	−0.0	−0.1	−0.1
S5	0.3	−0.2	0.1	0.5	0.2	0.2	−0.0	−0.1	−0.2	−0.6
S6	0.3	−0.2	0.2	0.5	0.2	−0.1	−0.0	0.1	0.2	0.6
S7	0.3	0.3	−0.5	−0.0	0.2	0.3	0.7	0.0	−0.0	0.0
S8	0.3	0.3	−0.5	0.1	0.2	−0.2	−0.7	−0.0	−0.0	−0.0
S9	0.2	0.5	0.4	0.0	−0.1	0.0	−0.0	0.7	−0.0	−0.1
S10	0.2	0.5	0.4	0.0	−0.1	0.0	0.0	−0.7	0.0	0.0
Percentage of total variation explained	52.3	20.4	11.0	8.5	5.0	1.0	0.9	0.6	0.2	0.2

The first component is of the type anticipated above, in this case accounting for 52.3% of the total variance. The second component contrasts arm measurements with leg measurements. Because of the overall positive correlations, individuals with strong reflexes at one site will tend to have strong reflexes at all other sites. What the second component does is to identify individuals with strong arm reflexes relative to their leg reflexes and contrast them with those individuals whose arm reflexes are weak compared to their leg reflexes. This component describes 20.4% of the total variance. The third component is a contrast of the ankle and wrist measurements with the others, which are further away from the ends of the limbs. The percentage of total variance explained is 11%. We leave the interpretation of the remainder of the components as an exercise for the reader. A similar example, in which all components have a clear simple interpretation, is given by Krzanowski[5] (p. 70). In that case the individuals are chickens, but once again body measurements (lengths of bones in this case) are being taken.

The α_{kj}s and λ_ks allow us respectively to interpret, and to measure the importance of, each PC, but the central part of any PCA is the set of PC scores. These can be used in a variety of ways. For example, if most of the variation in the data is accounted for by two PCs, a plot of the scores on these first two PCs will give an informative graphical display of the data. Various aspects of the data may come to light, such as clusters of patients or unusual patients, which are not obvious from the original data set, or from plots of any two of its variables. Such plots can be further enhanced by simultaneously plotting vectors of coefficients $\boldsymbol{\alpha}_k$, giving the so-called biplot. A number of examples of biplots in a medical context are given by Gabriel and Odoroff.[6] Although biplots arise naturally as a byproduct of PCA, these authors claim that they can be used as an informative graphical display of a data set, without any direct reference to PCA. There are similarities between biplots and correspondence analysis, a topic which is covered in a separate paper in this issue.

In Figure 1 we present the results of plotting the scores on the first two principal components for the data of Example 1. We shall leave detailed discussion of the features of this plot until Section 3. Just 43% of the total variance is explained by these two components.

PC scores can also be used as simple, lower dimensional, replacements for scores on the original variables, in some subsequent analysis of the data, for example in regression, discriminant analysis or cluster analysis. We comment on this further in Section 6.

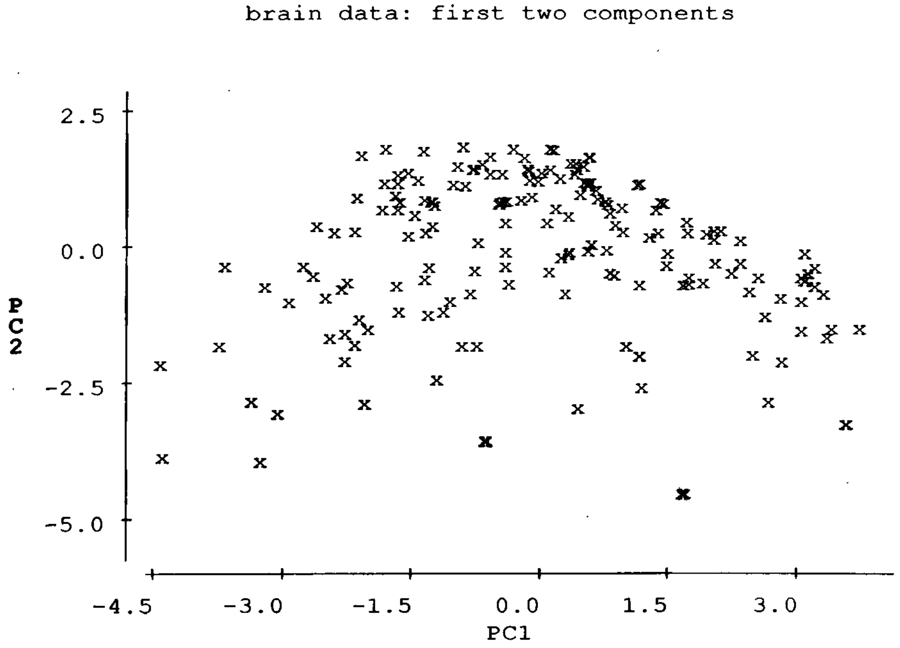

Figure 1 Plot of principal component scores for the first two components of the brain data.

2.2 Factor analysis

Like PCA, factor analysis attempts to reduce the dimensionality, p, of a data set consisting of p variables measured on n patients. The ways in which the two techniques are implemented are often similar, but the philosophy behind each is quite different. FA postulates that underlying the observed variables $x_1, x_2, \ldots, x_p$ there are m unobservable variables, or *common factors*, $y_1, y_2, \ldots, y_m$. The variables are assumed to be linear combinations of the factors plus, for each variable, an error term, or *specific factor*. Thus we have, for $j = 1, 2, \ldots,$

$$x_j = \sum_{k=1}^{m} \lambda_{jk} y_k + e_j,$$

or in matrix terms,

$$\mathbf{x} = \Lambda\,\mathbf{y} + \mathbf{e}$$

As with PCA we assume that the xs have been standardized, and we lose no generality if we assume also that the ys are standardized to have zero mean and unit variance. It is also assumed, initially, that the ys are uncorrelated. The coefficient λ_{jk} is known as the *loading* of the jth variable on the kth common factor, and if $\Psi_j = \text{var}(e_j)$, then $1 - \Psi_j$ is the *communality* for the jth variable (i.e. the proportion of the variation in x_j which is common to other variables, through the common factors).

To implement FA for given m we therefore need to find loadings λ_{jk} and communalities Ψ_j which give a good fit to the basic factor model above. It turns out that any solution which is found is not unique. In particular, if we multiply the matrix of factor loadings Λ by any orthogonal matrix $\mathbf{T}$, the resulting solution is as good as the original one. This multiplication by an orthogonal matrix is equivalent to an orthogonal rotation of the axes, so it is usual to talk of an (orthogonal) rotation of the initial factor solution. This rotation step can be made more general by allowing oblique (non-orthogonal) rotation.

Thus, any factor analysis consists of two steps, namely finding an initial solution, and then rotating that solution. Ways of finding an initial solution are discussed in Appendix A.2, and in Section 2.3 where links with PCA are examined. The role of rotation is that of simplification, indeed it is usual to talk about 'rotation to simple structure'. The idea is that among all the equally-good solutions, we should choose one which is easiest to interpret. We would therefore like as many as possible of the loadings to be close to zero or to ±1, so that it is clear which factors influence each variable and which do not. A number of criteria are available which quantify 'simple structure' – see Appendix A.2. In many examples it has been found that, for fixed m and initial solution, different rotation criteria lead to similar final solutions, so that it is often not crucial exactly which criterion is used. We shall see later that this is the case for the brain data.

In PCA we noted that there were three aspects to the results of an analysis. We now try to find similar constructs in FA and, in doing so we shall reveal some differences between the two techniques.

i) We shall usually have some measure of how much of the total variation in a data set is accounted for by m factors taken together, corresponding to $\sum_{k=1}^{m} \lambda_k$ in PCA. However, it is less clear how to divide this variation between individual factors, because rotation redistributes total variation among factors. A difference from PCA is that if we change from m to $(m + 1)$ factors we can completely change the nature of all the individual factors. Thus choice of m is often more important than the choice of rotation method in determining what the factors look like. By contrast, in PCA changing from m to $(m + 1)$ simply adds an extra component to those that were already there. In FA, in addition to the overall variation accounted for by m factors, we also have communalities which tell us how much of the variation in each individual variable is explained. The size of individual communalities, as well as overall variation, will need to be considered in choosing m.

ii) The loadings λ_{jk} in FA look superficially similar to the coefficients α_{kj} in PCA. However, in PCA we express the PCs as linear functions of the xs, whereas in FA the xs are linear functions of the factors, *plus an error term*. Unless we assume that the data are from a multivariate normal distribution, there is no guarantee that the factors in FA can be expressed as linear functions of the original variables. Despite this, at the simplest level there is a similarity in interpretation between the λ_{jk} and α_{kj}. Both measure which variables are related or not related to each factor or component.

iii) As noted in (ii), it is not, in general, possible to express the factors as linear functions of variables, which means that factor scores analogous to PC scores, cannot be calculated exactly. What can be done is to *estimate* such scores, and a variety of methods for doing this are available (see, for example, Lawley and Maxwell,[7] Chapter 8). Once estimated, the factor scores can be used in any of the ways suggested for PC scores, but factor analysis is usually more concerned with interpreting the factors underlying a data set (ie. looking at the loadings) than with subsequent analyses using the factor scores.

The title of this paper uses the term 'exploratory' factor analysis. The word 'exploratory' in this context means that we go into the analysis with no preconceived ideas of what the factor structure should be, except that we hope it will be simple. 'Confirmatory' factor analysis, which will be dealt with in a later issue of this journal, starts with some prior knowledge of the structure, such as the position of zeros among the λ_{jk}s. This restricted structure can then be estimated and tested (confirmed) for the data set of interest.

2.3 Connections between factor analysis and principal component analysis

The last part of the previous section stressed some differences between FA and PCA, and there are other differences too, some of which are noted later. The simplest difference is that FA assumes a model, while PCA does not. It is nevertheless true that there are often strong links between the two techniques in the way that they are implemented. In some special cases, the results of PCA and FA are identical or guaranteed to be very similar – see Jolliffe[3] (p. 124). On a more practical level, computer programs for PCA often include the technique as a special case of FA. The reason for this is that one way of finding an initial solution in the first step of FA is based on the first m PCs. Although of dubious validity ([3] p. 121), this is possibly the most commonly used method for finding an initial factor solution, as evidenced by the recent examples in the next section. From a pragmatic point of view, using PCs as initial factors frequently gives results which are not dissimilar to more 'respectable' methods.

2.4 A brief history and some recent examples

PCA was first introduced in the form discussed above by Hotelling,[8] but it had already appeared in a different, geometric formulation in Pearson.[9] Factor analysis has its roots in psychology, and dates from Spearman.[10] In its earliest form, certain well-defined structures (factors) were postulated underlying the results of various intelligence tests. As time went on, the method and its assumptions became less rigid, and took on the nature of an exploratory technique. To quote Gifi[11] (p. 319) '. . . [psychological] theory was banished . . . and factor analysis was no longer psychology but statistics.' Recently interest has again increased in stipulating well-defined structures (confirmatory factor analysis), and the area of applications has widened, though mostly within the social sciences. The fields in which PCA has been applied are broader than for FA, including pure sciences and humanities, as well as the social sciences. Despite the fact that a vast body of literature on factor analysis has been built up in psychology and elsewhere, neither FA nor PCA have been widely used in medical research generally. The coverage of two recent books on statistics for the analysis of medical data illustrates this point quite well. Everitt[12], has a single chapter on PCA, but does not include factor analysis. Albert and Harris,[13] despite its title *Multivariate interpretation of clinical laboratory data* has nothing on either topic. As part of the preparation for this paper, a number of medical statisticians were consulted, and a comprehensive literature search was carried out, but neither produced much further material. There is no clear reason for this lack of interest in the techniques. Although some medical data sets are small, large data bases are quite common for many branches of medical research and would often benefit from a judicious reduction of dimensionality. PCA and FA are obvious techniques to achieve such a reduction. There is a sprinkling of applications in the medical literature, but as noted in the introduction, confusion sometimes exists over terminology. We now review briefly a few recent examples.

Example 3: Psychiatry questionnaire: the dissociative experiences scale

Ross *et al.*[14] presents a classic straightforward application of factor analysis in a psychiatric context. A 28-item questionnaire was administered to 1055 individuals. The factor analysis used PCA to find initial solutions, and retained three components which, between them, accounted for 47.6% of the total variation in the original 28 variables. These three components were then rotated using the varimax criterion (see Appendix A.2) and a remarkably simple structure was reported. No variable has a nontrivial loading on more than one factor, and each factor is interpreted in psychiatric terms using these loadings.

Example 4: Audiologic tests: a composite score for a clinical trial

Henderson *et al.*[15] considered a data set consisting of scores on 24 audiologic tests for 65 patients. The main objective appeared to be to find a composite score based on the 24 tests, which could then be used to assess overall improvements in patients as a result of certain clinical trial treatments. PCA is an obvious technique for constructing such a score, since the first PC gives the (linear) composite with maximum possible variation. The first PC accounts for 61% of the total variation, an impressive result showing the large amount of redundancy in the information supplied by the 24 tests. However, this dramatic reduction is not entirely what it seems. The authors have based their PCA on the covariance matrix, rather than the correlation matrix, with the result that the first PC is dominated by those test scores with the greatest range of variability. Another special feature of this example is that the data are measured at several time periods, and the main result is based on pooling over time periods. We shall return to the topic of pooling data over several groups in Section 5.4.

Example 5: Respiratory function

Yamamura *et al.*[16] use PCA on 11 routine respiratory function tests for 88 patients. The first two PCs, which account for 38% and 22% of the total variation, are interpreted, respectively, as an index of the expiratory function and an index for overinflation of the lung. As in Henderson *et al.*,[15] data are pooled over different groups, but an additional feature of the present example is a plot of the scores on the first two PCs. Points on these plots are distinguished according to the value of a variable, which was not included in the PCA, and a clear relationship between the first PC and this additional variable emerges.

There is some confusion over terminology. Although a PCA is done, the text refers to *factor* scores and *factor* loadings rather than PC scores and PC coefficients. Some standard computer packages confusingly label PCs as 'factors' in this way. In Ross *et al.*,[14] there is also potential for confusion. It is only from a table heading that it can be deduced that PCA followed by varimax rotation has been done. The text simply states that 'PCA' identified three factors.

Example 6: Distribution of minor and trace elements in the human brain

In our final example of this section,[17] the title of the article refers only to PCA, but in fact the paper includes varimax rotation, and a further step leading to something which the authors call varimax rotated *absolute* principal component analysis. Their data consist of measurements of eight trace element concentrations in 46 human brain structures. In a PCA of these eight variables, plus one additional derived variable, the first two components accounted for 84% of the total variation. A plot of the scores on these two PCs reveals groups of similar brain structures.

The four examples of this section come from different fields, analysing data on psychiatric tests, audiologic measurements, respiratory function and brain structure. Everitt[12] gives another psychiatric example, and one from dentistry. (In the latter case, however, the analysis presented appears to be of the covariance matrix, while discussion is in terms of the correlation matrix.) The conclusion from these and the other examples in this paper is that in any area of medical research where large data sets are collected, PCA and FA have a potentially useful role to play.

3 A case study: extradural haematoma

3.1 The haematoma data

The principal component analysis of the diabetes patient data of Example 2 was remarkably straightforward. However, in that case there was a clear structure to the set of variables and this is somewhat unusual. In this section we provide detailed analyses of the somewhat unstructured data set of Example 1. As we can see from Table 4, eigenvalues now reduce gradually, rather than drop dramatically. Nevertheless, the first two components still describe 42% of the total variance of the 12 original variables. Coefficients for selected principal components are given in Table 5.

Table 4 Eigenvalues of correlation matrices: brain data

(i) *Using all 12 variables*

Eigenvalue	Percentage of variance	Cumulative percentage of variance
3.26	27	27
1.84	15	42
1.27	11	53
1.13	9	62
0.93	8	70
0.87	7	77
0.77	6	83
0.66	6	89
0.56	5	94
0.37	3	97
0.21	2	99
0.15	1	100

(ii) *Omitting the outcome variable*

Eigenvalue	Percentage of variance	Cumulative percentage of variance
2.88	26	26
1.83	17	43
1.13	10	53
1.11	10	63
0.93	8	71
0.87	8	79
0.77	7	86
0.66	6	92
0.45	4	96
0.24	2	98
0.15	1	99

(iii) *Omitting the outcome variable. Analysis for individuals with outcomes 1 and 2 (full recovery or minor disability)*

Eigenvalue	Percentage of variance	Cumulative percentage of variance
2.89	26	26
1.83	17	43
1.36	12	55
1.26	12	67
0.90	8	75
0.75	7	82
0.74	7	89
0.58	5	94
0.33	3	97
0.19	2	99
0.17	1	100

(iv) *As for (iii), but for individuals with poor outcomes (3–6)*

Eigenvalue	Percentage of variance	Cumulative percentage of variance
2.16	20	20
2.06	19	39
1.26	11	50
1.12	10	60
1.02	9	69
0.84	8	77
0.80	7	84
0.61	6	90
0.57	5	95
0.44	4	99
0.13	1	100

At one extreme of the first principal component we find individuals with large time intervals (and so presumably slow development of the haematoma), no deep coma and good outcomes. At the other extreme are individuals with small intervals (and presumably fast development of the haematoma), deep coma and poor outcomes. The negative coefficient for age is to be expected, as age is correlated with outcome, younger individuals with more flexible skulls being expected to be able to accommodate a swelling haematoma more easily without brain damage. The second principal component separates out individuals according to change in coma state between hospital arrival and prior to operation. With only a small number of exceptions, the coma-state measures deteriorate between the two measurements. High coma deterioration makes this component large, as does a small interval and the presence of alcohol. In a similar way we can interpret components 3 and 4, though proximity of the third and fourth eigenvalues may suggest that rotation of these components, which do not have such a clear interpretation as the first two, will clarify the picture (see Section 5.1). The last

Table 5 Coefficients for the first four and the last two principal components, for the brain data, based on correlation matrices, (i) using all 12 variables, (ii) omitting the outcome variable, (iiii) omitting the outcome variable: analysis for individuals with outcomes 1 and 2, (iv) as for (iii) but for individuals with poor outcomes. In all cases we just give 10 × coefficients rounded to one decimal place

	(i)						(ii)					
	Principal component						Principal component					
Variable	1	2	3	4	11	12	1	2	3	4	10	11
Age	−2	−3	−6	1	−1	0	−1	2	3	6	0	0
Sex	0	1	−2	−6	0	0	0	−1	6	−3	0	0
Interval	4	−3	−2	0	0	0	4	3	0	0	−1	0
Tests	2	−3	−2	0	0	0	3	2	1	2	0	0
Alcohol	−1	−5	1	2	0	0	−1	5	−2	0	0	0
Anaesthesia	2	−1	3	−5	0	0	2	1	2	−7	1	0
VADM	−3	−5	2	−2	2	7	−3	5	1	−2	−2	7
VP	−4	3	0	−1	6	−2	−5	−3	1	0	−7	−2
MADM	−3	−4	2	−2	−1	−7	−4	5	1	−2	1	−7
MP	−5	2	0	−1	−7	2	−5	−2	1	0	7	2
Distance	1	0	−5	−5	0	0	1	0	7	1	0	0
Outcome	−4	−1	−4	0	3	0						

	(iii)						(iv)					
	Principal component						Principal component					
Variable	1	2	3	4	11	12	1	2	3	4	10	11
Age	2	−2	−4	−4	−1	0	0	−1	7	4	1	0
Sex	0	2	6	0	−1	0	1	0	6	−3	0	1
Interval	5	−1	2	1	0	0	4	−3	−1	1	0	0
Tests	3	−1	−1	−3	0	0	1	−2	2	−6	1	−1
Alcohol	1	−3	−3	0	0	0	−2	−4	0	−2	1	−1
Anaesthesia	1	1	4	−4	0	0	1	−3	−4	−4	1	0
VADM	−2	−6	3	−1	6	4	−5	−4	−1	1	−1	7
VP	−5	2	−1	−3	5	−5	−3	4	0	−2	−6	−2
MADM	−2	−6	2	0	−5	−5	−5	−3	0	2	1	−7
MP	−5	0	0	−3	−5	5	−3	5	0	−1	7	1
Distance	2	0	1	−7	0	−1	2	0	−2	4	1	0

two components correspond to two simple contrasts of the coma scores which have small eigenvalues (variances).

The outcome variable is qualitatively different from the others, but when it is omitted the changes are changes of degree rather than dramatic changes in interpretation. (See Tables 4 and 5.) Age is correlated with outcome and it is not surprising that its role in components 3 and 4 is somewhat changed when outcome is omitted. However apart from overall changes in sign, which have no effect within any component, the components appear to be roughly as before but in the different order when outcome is not included.

Figure 1 (see Section 2.1) reveals that high deterioration of the coma variables between the two occasions of measurement is far more common than otherwise. The plot of points is suggestive of two subgroups, corresponding roughly to good and poor outcomes. In Tables 4 and 5 we provide the results from having performed separate principal component analyses based on one division of individuals into such outcome categories. The eigenvalue distribution is fairly similar to before except that the first component for the poor outcome individuals is now less dominant. The pattern of signs for components 1, 2, 10 and 11 is effectively unchanged from before with the exception of the coefficients for the distance variable. As might be expected, the greatest changes occur for components 3 and 4, where interpretation may, again, be simplified by rotation.

This leads us naturally to consider the results from factor analysis. We have, for illustration, considered two initial configurations, one being the principal component solution and the other being the maximum-likelihood solution. Both of these are then rotated by means of both varimax and quartimax methods separately. We shall give the results for the first four factors in each case, starting with the principal component initial solution. For simplicity of comparison we shall just consider the case of all 12 variables taken together.

For the principal component starting configuration, the varimax and quartimax solutions are strikingly similar, as seen from Table 6. The rotations have clearly assisted with the interpretation of factors 3 and 4, but a complete intuitive explanation is still difficult to provide. Factor 3 emphasizes the relationship between age and outcome. Rotation to simple structure means that certain contrasts before rotation may disappear following rotation, and we can see that this has occurred in factor 2. Factors 1 and 2 now relate to outcome and coma scores; factor 1 places weight on the scores prior to the operation, while factor 2 places weight on the score on hospital admission. Certain other variables also have important coefficients as can be seen.

In Table 7 we give the factor coefficients based on the maximum likelihood approach, as well as those obtained following varimax and quartimax rotations. A striking feature is that the four factors now only describe 42% of the total variance, compared with 62% above. This is because, unlike PCA, the maximum likelihood factors are not designed to maximize variance. The interpretation of the first factor is as before, now with greater weight being placed on outcome. Factor 2 is qualitatively similar to the above rotated factor 2 (principal component start), placing weight on the coma variables prior to the operation, while factor 3 places weight on the coma variables at admission to hospital. Factor 4 places most weight on distance and anaesthesia. The rotated factors using the varimax procedure are remarkably similar to the rotated factors using the quartimax

Table 6 Factor loadings for the brain data for the first four factors, (i) varimax solution and (ii) quartimax solution, in each case starting from the principal component solution described in Tables 4(i) and 5(i). We give 10 × loadings rounded to one decimal place.

		Factor			
	Variable	1	2	3	4
	Age	−1	2	8	1
	Sex	1	−0	−1	7
	Interval	−8	−0	−0	1
(i)	Tests	−5	0	1	1
	Alcohol	−3	5	1	−3
	Anaesthesia	−2	2	−6	4
	VADM	1	9	1	0
	VP	8	1	2	1
	MADM	2	9	1	0
	MP	8	1	3	4
	Distance	−2	−1	2	7
	Outcome	4	3	7	1
	Variance	2.7	2.1	1.6	1.2

		Factor			
	Variable	1	2	3	4
	Age	−1	2	8	1
	Sex	0	−0	−1	7
	Interval	−8	−0	−0	1
(ii)	Tests	−5	0	1	1
	Alcohol	−2	6	1	−3
	Anaesthesia	−2	2	−6	4
	VADM	2	9	1	0
	VP	8	1	1	1
	MADM	3	9	1	0
	MP	8	1	2	1
	Distance	−2	−1	2	7
	Outcome	4	3	7	1
	Variance	2.7	2.1	1.5	1.2

Table 7 Factor loadings for the brain data for the first four factors, (i) maximum likelihood starting solution, (ii) subsequent varimax solution and (iii) subsequent quartimax solution. We give 10 × loadings rounded to one decimal place.

		Factor			
	Variable	1	2	3	4
	Age	3	−3	0	0
	Sex	−0	−0	−1	2
	Interval	−5	−5	−0	1
(i)	Tests	−2	−3	−0	1
	Alcohol	1	−2	3	−0
	Anaesthesia	−3	0	0	2
	VADM	3	−0	6	1
	VP	6	6	−1	1
	MADM	4	−0	7	1
	MP	8	5	−0	0
	Distance	−0	−1	−2	5
	Outcome	9	−3	−1	−0
	Variance	2.6	1.1	1.0	0.3

		Factor			
	Variable	1	2	3	4
	Age	−0	2	4	1
	Sex	−0	−0	−0	2
	Interval	−7	−0	−0	2
(ii)	Tests	−4	−0	−0	1
	Alcohol	−1	3	1	−1
	Anaesthesia	−2	0	−3	2
	VADM	2	7	1	−0
	VP	9	0	1	1
	MADM	2	8	1	−1
	MP	8	1	3	0
	Distance	−1	−1	0	5
	Outcome	3	2	9	1
	Variance	2.3	1.3	1.2	0.4

		Factor			
	Variable	1	2	3	4
	Age	−0	2	4	1
	Sex	−0	−0	−0	2
	Interval	−7	−0	−0	1
(i)	Tests	−4	−0	0	1
	Alcohol	−1	3	1	−1
	Anaesthesia	−2	0	−3	2
	VADM	2	7	1	−0
	VP	9	0	1	1
	MADM	2	8	1	−1
	MP	8	1	2	1
	Distance	−1	−1	1	5
	Outcome	4	2	9	1
	Variance	2.4	1.3	1.1	0.4

procedure. Factors 1 and 2 are very similar to rotated factors 1 and 2 following the PCA solution. Factor 3 is quite similar to rotated factor 3 following the PCA solution, but Factor 4 is different (compared with rotated factor 4 following the PCA solution), being little changed in comparison with the unrotated factor 4 for the maximum-likelihood solution.

In summary, the principal component analysis requires a fair number of components to describe the data. The first two components have an interpretation in terms of quality of outcome/coma state and rate of decline into coma. Other components are less easy to interpret. Factor analyses, following rotation, preserve the basic description of the first principal component, but lose the interpretation of the second principal component, which involved a contrast of coma scores taken over time. Later factors are to some extent easier to describe than later components, but a clear intuitive explanation is not forthcoming. Two avenues for future research are further analyses based on contrasts of coma variables as well as the original variables themselves, and rotation of just the third and fourth components, a topic to which we shall return in Section 5.1.

3.2 Additional features/examples

In discussing the principal component analysis above we analysed two subsets of the data as well as the entire data set. The differences which arose for the subsets were relatively minor, and readily explained. The presence of striking heterogeneity in a set of data can, however, result in quite striking differences when subgroup analyses take place. We now provide a (nonmedical) example of this.

Example 7: Iris data

The famous Fisher Iris data involve four length measurements taken on each of 150 Irises, divided equally between three different species (Andrews and Herzberg,[18] p. 5). The measurements were of sepal length and width and petal length and width. A principal component analysis of the correlation matrix produces the results of Table 8.

Table 8 The results of a principal component analysis of the Fisher iris data: all 150 irises, irrespective of species; coefficients and eigenvalues.

	Principal component			
Variable	1	2	3	4
sepal length	−0.52	0.38	0.72	0.26
sepal width	0.27	0.92	−0.24	−0.12
petal length	−0.58	0.02	−0.14	−0.80
petal width	−0.56	0.07	−0.63	0.52
eigenvalues	2.92	0.91	0.15	0.02

It was commented in the discussion of Example 2 that the first component, when data involve length measurements, is often an overall measure of size, but that does not occur here. However, if the species are analysed separately, the first principal component is, in each case, a weighted average of the four measures (i.e. a measure of size) explaining, for the three species in order, 52%, 73% and 61% of the total variance. The full results for all three species separately are shown in Table 9, and we can also see similarities of interpretation of the later components, between species.

The key point here is that the first component for the full data set separates out the different species, and the direction of the first component is different from those of the within-species data sets. For any data set the possibility of heterogeneity must be borne in mind. In both this example, and Figure 1, we are alerted to the possibility of heterogeneity from plots of principal component scores.

The comparison of principal components, from the three different species, may be formalized using the approach of Krzanowski.[19] The particular comparisons of Table 9 might be facilitated by judicious axis rotation (see below). Further discussion of PCA in the presence of groups of observations is given in Section 5.4.

Classic examples of principal component interpretation are found in Jeffers[20] and Moser and Scott.[21] A further complex example is discussed in detail by Jolliffe, Jones and Morgan,[22] where many variables were measured on 2622 elderly individuals.

Table 9 Principal component analysis of the Fisher iris data: results as for Table 8, but now for each species separately

Species setosa	Principal component			
Variable	1	2	3	4
sepal length	0.60	−0.34	0.07	0.74
sepal width	0.58	−0.44	0.00	−0.69
petal length	0.38	0.63	0.68	−0.09
petal width	0.41	0.55	−0.73	−0.02
eigenvalues	2.06	1.02	0.67	0.25

Species versicolor	Principal component			
Variable	1	2	3	4
sepal length	−0.48	−0.61	0.49	−0.39
sepal width	−0.47	0.67	0.54	0.20
petal length	−0.54	−0.31	−0.34	0.71
petal width	−0.52	0.28	−0.59	−0.55
eigenvalues	2.93	0.55	0.40	0.13

Species virginica	Principal component			
Variable	1	2	3	4
sepal length	0.55	−0.43	−0.01	0.71
sepal width	0.48	0.44	−0.75	−0.12
petal length	0.55	−0.43	0.20	−0.69
petal width	0.41	0.66	0.63	0.10
eigenvalues	2.45	0.96	0.45	0.13

4 Further details of implementation

4.1 How many components or factors?

In the examples discussed so far we have talked mainly in terms of retaining sufficient components or factors to achieve a high percentage of the original variation. In factor analysis communalities also play a role in the choice of m, the number of factors to retain. It might be asked whether we can improve on this rule-of-thumb. We shall discuss here one other, equally informal, way of choosing m when the analysis is based on a correlation matrix. Many other methods for choosing m have been proposed. Jolliffe[3] (Section 6.1) gives a partial review in the context of PCA and concludes that '. . . rules which have more sound statistical foundations seem, at present, to offer little advantage over the simpler rules'. Despite further attempts by various authors since 1986 to produce improved rules, the conclusion remains valid. Indeed, some recent work has cast doubt on the worth of the cross-validation rules, described in Section 6.1.5 of Jolliffe,[3] which had originally appeared promising.

A second simple rule for PCA is to retain all those components whose eigenvalues are greater than some cutoff. The same rule can be used for FA, if PCA is used as an initial solution, but the optimal level of the cutoff should be different for the two techniques. The most usual cutoff is the average value of all eigenvalues. It is simplest to discuss this for correlation matrices, for which of course this average is unity, or equivalently the contribution of one of the original standardized variables to the total variation. It can be argued that the cutoff should be lower than this for PCA but higher for FA, for the following reason. Suppose that one of the original variables is almost independent of all other variables. One of the PCs will then be dominated by this variable, with a variance close to unity, although sampling variation will mean that the variance could be slightly higher or lower. In PCA we want to ensure that we do not throw away information which cannot be reproduced by other components, so we shall keep such one-variable PCs. To make sure they are kept, and to allow for sampling variation, the cutoff should be less than 1. Jolliffe[23] suggested a cutoff of 0.7, although this is not a hard-and-fast rule.

In the context of FA, one-variable factors are not of interest. They are specific factors whereas we are looking for common factors. The cutoff in FA should therefore be set higher than 1, to ensure that we exclude such factors.

4.2 Types of data

Although multivariate normality has been mentioned earlier in passing, we have deliberately avoided making any distributional assumptions. It was clear in the main example of Section 3 that some variables were continuous, some had ordered categories, some were dichotomous and so on. It may have seemed somewhat cavalier to throw such a mixture of variables into a PCA, but we would argue that it is acceptable, subject to a certain amount of caution in the interpretation of results. Admittedly, there is a great deal of theoretical literature on PCA based on the multivariate normal distribution, and the same is true to a lesser extent for FA. If it is required to test formal hypotheses, construct confidence intervals or carry out other inferential procedures, then we will need some of this distributional edifice. However, the remit of this paper, and indeed the major role of PCA and FA in practice, is exploratory. In this case there can be no overriding objection to the presence of different types of variables, although we should, for example, be on the lookout for artifacts in plots of PC scores caused by discreteness in some of the variables. One illuminating result which supports the case for using PCA on discrete, or even binary, variables, is that PCA on binary variables is equivalent to the well-established technique of principal co-ordinate analysis (also known as metric scaling) using a standard similarity measure.[24]

Example 8: Anorexia nervosa

The data forming the basis for this study are described in the unpublished paper by Beumont, Booth, Abraham, Griffiths and Turner.[25] Twenty-five women suffering from the slimmers' disease, *Anorexia nervosa*, were questioned on the nature of their symptoms. Here we record whether, for each woman, a symptom was present or not. The incidence matrix is shown in Table 10. Only two symptoms (4: amenorrhoea and 20: manipulating body fluids) are present in all of the women.

If we perform a principal component analysis on the data of Table 10, based on the *covariance* matrix, the end-result is a metric scaling of the simple matching coefficients. For these data we present the resulting plot in Figure 2. In Figure 2(a) we have added a minimum spanning tree obtained from the simple matching coefficients, which is a convenient device for investigating the distortion in the two-dimensional representation (Krzanowski,[5] p.103). Another approach to investigating distortion is to add to the plot the total numbers of symptoms present for each woman. This was suggested by Banfield and Gower[26] and is shown in Figure 2(b). Further analyses of this example are given in an unpublished paper by Morgan and Ray.[27] Both the simple matching coefficient and the minimum spanning tree are well-described in Digby and Kempton.[28]

Table 10 Not all women suffering from anorexia nervosa exhibit the same symptoms. For a set of 25 women (the columns of the table) this table describes their symptoms, out of a complete set of 28 given by the rows of the table. '1' indicates a symptom is present and '0' indicates the symptom is absent

0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	1	1	0	1	1	0
1	1	1	1	0	1	1	1	0	1	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1
0	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0	0	1	0	1	1	0	1	1	0	0	1	1	0	0	0	1	0	1	1	1	1	1	0	1
0	0	0	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	0	1	1	1	1	1	1
1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	0
0	0	0	0	1	0	0	1	1	1	1	1	0	0	0	0	0	1	0	1	1	1	1	0	1
1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	1	0	0	0	1	1	1
0	0	1	1	1	1	1	1	1	1	0	1	1	0	1	1	1	0	1	1	1	1	1	1	1
0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1
1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0	1	1	0	0	0	1	1	1	0	1	1	0	0	0	1	1	0	1	1	1	1	1	1	1
1	0	1	1	0	0	1	1	1	1	0	0	0	0	1	0	1	1	1	0	0	0	1	0	1
1	1	1	1	1	1	1	1	1	0	0	0	1	1	0	0	1	0	1	1	1	0	1	1	1
0	1	1	1	0	1	1	1	1	1	0	1	1	0	1	0	1	1	1	0	0	1	1	1	1
0	1	0	1	0	0	0	0	1	1	1	1	0	0	0	1	1	0	1	0	0	1	0	1	1
1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	0	1	0	1
1	0	0	0	0	1	0	1	1	1	1	1	1	1	0	0	1	1	1	0	1	1	1	1	1
0	0	1	0	1	0	0	1	1	1	0	1	1	0	1	1	1	0	0	0	1	1	1	0	1
0	0	0	1	1	0	0	0	1	0	1	1	1	0	0	1	0	0	0	1	1	0	1	1	1
0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0	0	1	0	1	0	1	0	1	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	1	0	0	1	0	1	0	0	1	0	0	1	1	1	1	0	1	0	0

4.3 Normalization constraints

In the introduction we noted that normalization constraints on the vectors of coefficients $\boldsymbol{\alpha}_k$, were needed to make our problem well-defined. We chose the constraint $\boldsymbol{\alpha}_k'\boldsymbol{\alpha}_k = 1$, but after deriving the PCs we are quite at liberty to change the constraint without changing the character of a PC. It is the relative, rather than absolute, values of α_{kj} within $\boldsymbol{\alpha}_k$ which determine the interpretation of the corresponding PC. For example, if $\boldsymbol{\alpha}'\alpha = 1$, then setting $\boldsymbol{\gamma} = (\lambda^{1/2}\boldsymbol{\alpha})$ results in $\boldsymbol{\gamma}'\boldsymbol{\gamma} = \lambda$. This is one of two alternative normalizations that are of particular interest. If we require that $\boldsymbol{\alpha}_k'\boldsymbol{\alpha}_k = \lambda_{kj}$, then when PCA is based on correlation matrices, the α_{kj} are the correlations between the kth PC and the jth variable. This normalization is also analogous to what is usually done in FA. There the assumption of unit variances for both the variables and the factors, coupled with independence of the factors, means that λ_{kj} is the correlation between the jth variable and the kth factor.

Another possible normalization for PCs is $\boldsymbol{\alpha}_k'\boldsymbol{\alpha}_k = 1/\lambda_k$, which leads to $\text{var}(z_k) = 1$ for all k. This normalization is useful in outlier detection as we shall note in Section 5.2.

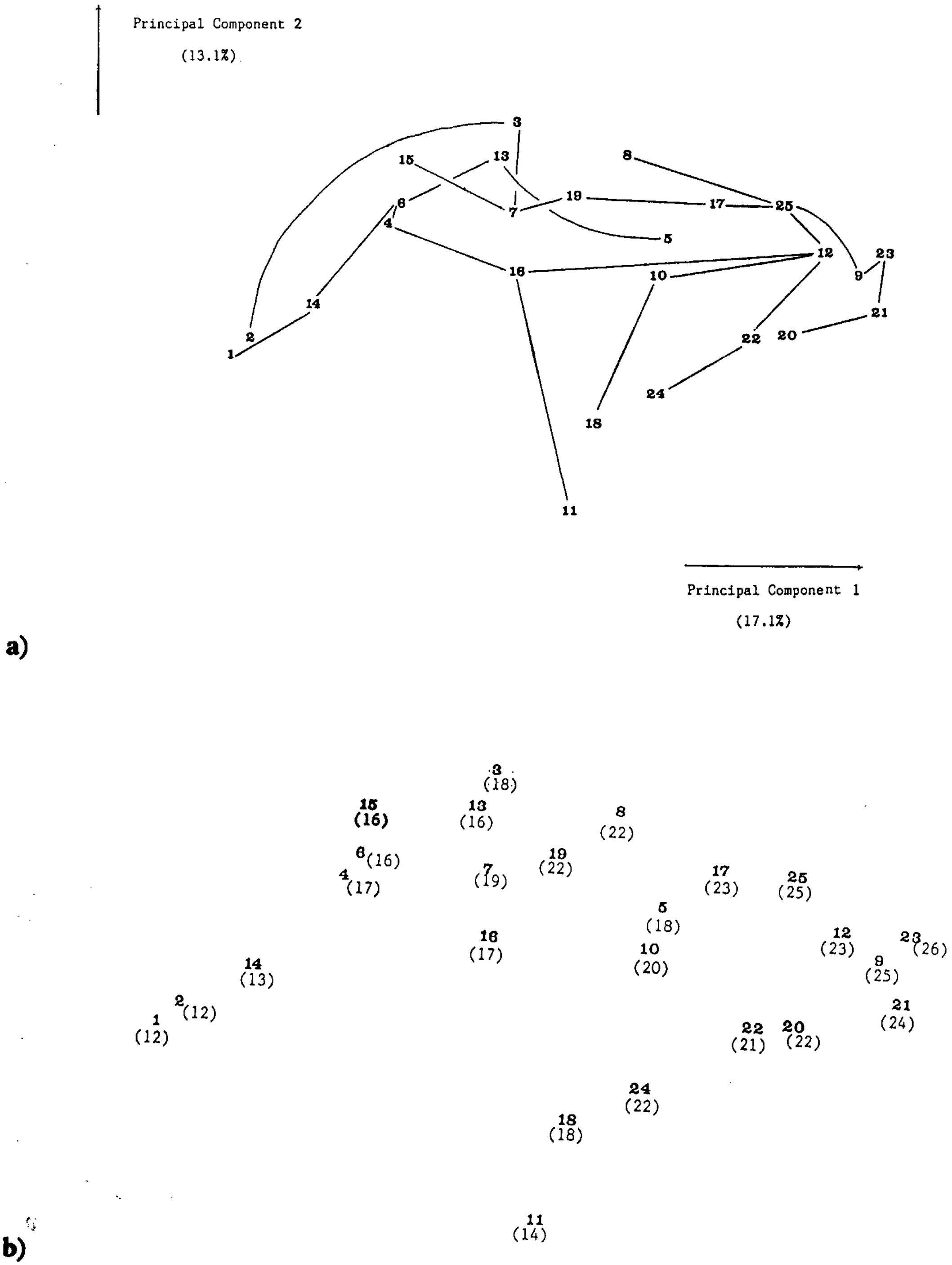

Figure 2 Plot of the 25 anorexic women, using the scores on the first two principal components: (a) with minimum spanning tree from the simple matching coefficients; (b) with total number of symptoms present for each woman, added in parentheses.

5 Modifications and extensions

Apart from Section 5.1, the remainder of this section concentrates on PCA rather than FA. One reason for this emphasis is that confirmatory factor analysis, which is a major extension to what we have done, is to be covered in an article in a subsequent issue of the journal.

5.1 Rotating principal components

We have noted that the version of FA which is most often implemented, simply rotates the first few PCs. We have also noted that, in the context of the factor model, this is usually a rather dubious practice. However, there is nothing inherently wrong with rotating PCs in an attempt to simplify their structure, and we could do so without calling the resulting technique 'factor analysis'.

A problem with rotating the first few components is that although the total variation accounted for by the rotated components is the same as before rotation, it is more evenly distributed after rotation. We may therefore lose sight of well-defined major sources of variation. It is well-known that PCs whose eigenvalues are well-separated from those of adjacent PCs are stable and well-defined. PCs whose eigenvalues are close together are, however, unstable and ill-defined, although the space that a group of such PCs span together is stable if the group has eigenvalues well-separated from those of adjacent PCs. This instability manifests itself in potentially large changes in individual PCs caused by small changes in the data set. These changes are, in fact, rotations within the stable space spanned by the group of PCs. It therefore seems sensible to rotate groups of individually ill-defined PCs to achieve the simplest possible structure within their subspace while leaving unrotated any well-defined individual PCs. This idea is illustrated by an example from Jolliffe.[29]

Example 9: Facial spots

The data set consists of five variables measuring the number of facial spots in five categories of severity, for 34 individuals. In a PCA based on the correlation matrix the first three eigenvalues are 2.70, 0.87 and 0.67. The first PC has positive coefficients on all five variables and can be interpreted as an overall measure of spottiness. The coefficients of the second and third PCs are given in Table 11, and it is seen that the second is dominated by the fifth variable, whereas component 3 is a more complicated contrast between variables, with variables 1 and 3 most important.

The plot of the PC scores shown in Figure 3 shows two clear outliers. When the outlier on the second PC is omitted, Components 2 and 3 appear to change drastically (see Table 11 – Component 1 is virtually unchanged).

Table 11 Facial spots data: coefficients of components 2 and 3

Unrotated				Rotated	
Outlier included		Outlier omitted		Outlier omitted	
Component 2	Component 3	Component 2	Component 3	Component 2	Component 3
−0.23	0.69	0.53	−0.49	0.69	−0.20
−0.22	0.22	−0.03	−0.49	0.19	−0.46
0.01	−0.56	−0.51	0.13	−0.52	−0.12
−0.12	−0.36	−0.36	0.22	−0.42	0.03
0.94	0.18	0.57	0.68	0.20	0.86

However, the first three eigenvalues are now 2.79, 0.74, 0.70 and the closeness of eigenvalues 2 or 3 suggests instability of their individual eigenvectors. Rotation of these two eigenvectors using the Varimax method gives results as in Table 11. We see from the table that, apart from a switch in order, we are almost back to the original components, before removing the outlier. Thus the effect of omitting the outlier is much less than appeared at first. Rather than drastically changing the second and third PCs it has simply made them less stable. Rotation of the unstable pair has recovered the original structure, though it is somewhat fortuitous that the original structure was 'simple' and hence found, approximately, by simple structure rotation.

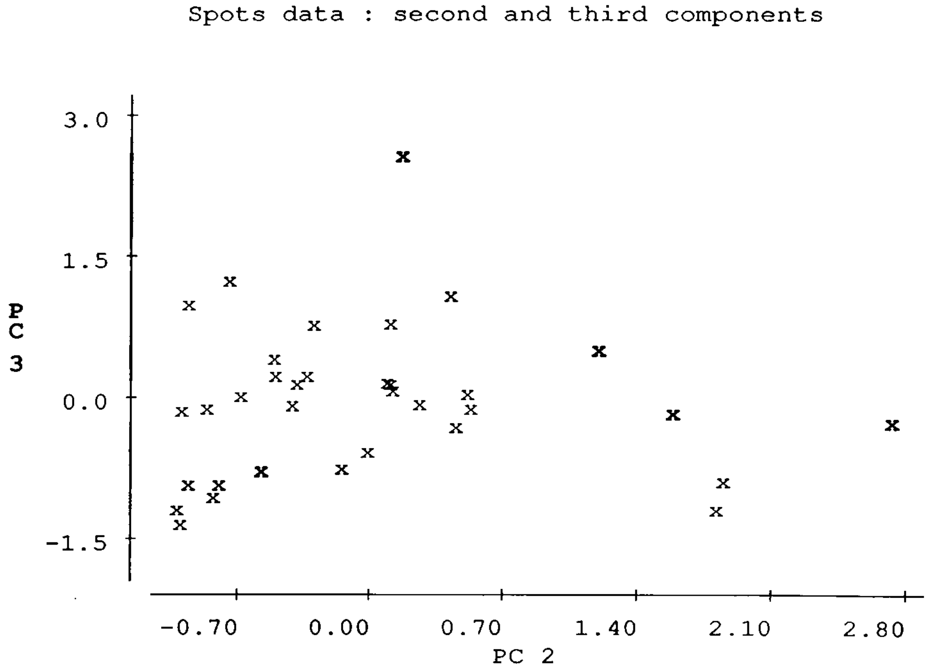

Figure 3 Plot of principal component scores for the second and third components of the facial spots data

5.2 Looking for structure in the data

The heading above refers to structure among the observations, rather than among the variables. It could be argued that the main objective of PCA and FA is looking at the structure, or associations between variables, but by plotting Factor or PC scores we may also find structure amongst observations. For example, Duflou *et al.*[17] (Example 6) found groups of observations in their plot of PC scores, and the facial spots example (Example 9) in the previous section identified two outliers in the same way. Clusters of observations, and outliers are the main types of structure we might hope to find, but there is, of course, no guarantee that such features will show up on the first two or three PCs or factors. In fact, it can be demonstrated that the last few PCs, which are often discarded as noise, may be better at detecting certain types of outlier than are the first few PCs – see Jolliffe ([3], Section 10.1), especially if we use the alternative normalization $\alpha_k'\alpha_k = 1/\lambda_k$ when combining components. We repeat, however, that PCA and FA are not *designed* to find outliers or clusters, and if that is our main objective in dimension reduction, we should use a purpose-built technique such as projection pursuit – see, for example Huber,[30] Jones and Sibson,[31] Friedman.[32] Recently in an unpublished dissertation Morton[33] incorporated the idea of simple structure via rotation into the projection pursuit paradigm.

5.3 Influence of individual observations

Even in quite large samples, clear outliers can exert an appreciable effect on a PCA or FA (Pack *et al.*).[34] It can also happen that less extreme individual observations have a disproportionate effect on the results of a PCA or FA. Although there is little chance of this in large samples and for PCs with well-spaced eigenvalues, it is certainly a possibility for small samples, and for PCs with nearly equal eigenvalues. We have already seen an example of this phenomenon in Section 5.1. Work has been done in the past few years on how to assess the likely influence of individual observations, both in PCA and in certain types of FA – see, for example, Critchley,[35] Pack *et al.*[34] and Tanaka and Odaka.[36]

5.4 Common principal components, repeated measures

In two of the examples discussed in Section 2.4 (Examples 4 and 5), the observations were in a number of groups, and the analysis was done on data pooled across all groups. Group structure was also suggested in the main example of Section 3 (Example 1) and was certainly present in Example 7. Before any pooling across groups can take place, we need to be reasonably confident that the covariance or correlation matrices in the different groups are similar enough to be combined, i.e. that the sample correlation or covariance matrices have arisen from samples from the same population. If not, then we could do a separate analysis for each group, but in some cases an intermediate approach may be sensible. It may be possible to assume some relationship between covariance matrices in different groups, without assuming identity of these matrices in their respective populations. Flury[37] gives a hierarchy of models which satisfy these conditions. His results are largely for covariance, rather than correlation, matrices, and assume independence of samples for different groups. Of course a correlation matrix is also a covariance matrix, but distributional results are easier to derive in the latter case than in the former case because a correlation involves *ratios* of random variables. Independence is a reasonable assumption where different groups consist of different patients, as in Yamamura *et al.*[16] However, when the different groups are the same patients measured at different times as in Henderson *et al.*[15] (the repeated measures case), independence is at best a dubious assumption. A recent, as yet unpublished, paper by Flury and Neuenschwander[38] extends Flury's models to dependent data, and could have great potential for describing repeated measures data.

5.5 Nonlinear PCA

One great convenience of both PCA and FA is their linearity. It keeps things simple, but at the same time is rather restrictive. A more flexible approach is to use non linear functions, and there are clearly a number of ways in which this might be done. One of the most comprehensive references for nonlinear multivariate analysis is Gifi.[11] The general idea behind Gifi's approach is discretization of all variables, followed by an optimal nonlinear transformation of category scores, before carrying out a PCA or some other optimization procedure, on these transformed variables. Although Gifi's examples are largely from social science, and not from medicine, some of the case studies in the final chapter of Gifi[11] illustrate well the extra flexibility obtained by allowing nonlinearity.

6 Discussion

The view is advanced in Chatfield and Collins,[39] that factor analysis 'should not be used in most practical situations'. These authors list six shortcomings of factor analysis. While we do not take such an extreme view in this paper, we have focused on principal components and have shown how rotation of components can, in some cases, be useful without having to consider a full-blown factor analysis.

When principal component analysis is successful in greatly reducing the dimensionality of a problem, that is clearly a major step forward in understanding and describing the data. It must be realized that each component is still a combination of all of the original variables, but one may be fortunate enough to obtain a simple interpretation of components. Sometimes this suggests a similar description of the data through using a subset of suitably chosen variables. For an investigation of this see Jolliffe.[40]

As mentioned already, principal components may be used in a range of further statistical analyses. For a recent application in generalized linear regression see the paper by Marx and Smith.[41] Principal components can even be a useful preliminary step in 'alternative' dimension-reducing techniques such as projection pursuit[32] and Andrews' curves.[42] However, some caution is necessary when choosing which PCs should go into a subsequent analysis such as regression, discrimination or cluster analysis. For example, when principal components are used in regression or discrimination it is not necessarily always the components with the largest variance which provide the best prediction. For discussion see Jolliffe,[3] Sections 8.2, 9.1 and 9.2, and Chang.[43]

Postscript

We have discussed above the possible influence that individuals may exert on the outcome of a principal component analysis. It is also important to consider carefully before any analysis of which variables to include, whether they have been coded/scaled correctly, etc. It has probably escaped notice, for instance, that one of the variables in the brain data, namely the 'tests' variable is not ordinal, and due account has not been taken of this. In fact, omission of that variable does not affect the conclusions drawn. The full brain data set is available through electronic mail and the JANET network.

Appendices

A1: Derivation of principal components

Suppose that $\mathbf{x}$ consists of p variables, standardized so that each has unit variance. The first principal component is defined as the linear function $z_1 = \boldsymbol{\alpha}'\mathbf{x}$, where the vector of coefficients, $\boldsymbol{\alpha}$, is chosen so that $\text{var}(z_1)$ is maximized, subject to the normalization constraint $\boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$. It is easy to verify that $\text{var}(z_1) = \boldsymbol{\alpha}'\mathbf{R}\boldsymbol{\alpha}$, where $\mathbf{R}$ is the correlation matrix for $\mathbf{x}$, so we maximize.

$$\boldsymbol{\alpha}'\mathbf{R}\boldsymbol{\alpha} - \lambda(\boldsymbol{\alpha}'\boldsymbol{\alpha} - 1)$$

where λ is a Lagrange multiplier. Differentiating with respect to $\boldsymbol{\alpha}$ and equating the derivative to a vector of zeros leads to the equation $\mathbf{R}\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha}$. This is an eigenequation, so that λ is an eigenvalue of $\mathbf{R}$. Multiplication of the eigenequation on the left by α' shows that $\mathrm{var}(z_1) = \lambda$, and hence λ must be as large as possible. Thus $\mathrm{var}(z_1) = \lambda_1$, the largest eigenvalue of $\mathbf{R}$, and $\boldsymbol{\alpha}$ is the corresponding eigenvector.

In general, for the kth PC we maximize $\mathrm{var}(z_k) = \boldsymbol{\alpha}_k'\mathbf{R}\boldsymbol{\alpha}_k$ subject to $\boldsymbol{\alpha}_k'\boldsymbol{\alpha}_k = 1$, with the additional constraints $\boldsymbol{\alpha}_k'\mathbf{R}\boldsymbol{\alpha}_j = 0$ $(= \boldsymbol{\alpha}_k'\boldsymbol{\alpha}_j)$, $j = 1, 2, \ldots, k - 1$. Introducing a Lagrange multiplier for each constraint, we use a similar procedure to that for the first PC to show that $\mathrm{var}(z_k) = \lambda_k$, the kth largest eigenvalue of $\mathbf{R}$, and that $\boldsymbol{\alpha}_k$ is the corresponding eigenvector.

A2: Some possible methods of factor analysis

We have noted in Section 2.2 that factor analysis has two stages; first, we find an initial solution, and then we rotate to simple structure. There are many possibilites for each stage, and hence a large number of possible factor analysis procedures. Here we simply describe a few of the possibilities.

For the first stage, we have already mentioned that PCs are often used as an initial solution. A simple modification, called principal factor analysis, replaces the unit values on the diagonal of the correlation matrix by estimates of the communalities of each variable, and then carries out an eigenanalysis of this modified matrix. A statistically more respectable initial solution, which is more closely related to the factor model, is based on maximum likelihood estimation of the parameters in the model. Full details can be found in Lawley and Maxwell.[7] Of course, to obtain maximum likelihood estimators we need distributional assumptions, specifically multivariate normality. This may seem restrictive, but fortunately the estimators found this way also have a derivation in terms of partial correlations, which does not explicitly rely on any distributional assumptions – see Morrison,[44] Section 9.8. An additional advantage of the maximum likelihood method is that the results are the same whether we use a correlation or covariance matrix. This is in complete contrast to PCA, or FA when PCs are used as initial solutions, where the results obtained from covariance and correlation matrices can be quite different, and are not readily derived from each other.

Numerous methods have been suggested for rotating factors. Cattell[45] and Richman[46] give nonexhaustive lists of 11 and 19 possibilites respectively. Here we simply describe what is probably the most common criterion (it is the default in several computer packages), the varimax criterion. Even varimax comes in more than one version, but in the simplest we maximize the sum over factors of the within-factor variances of squared loadings,

$$\sum_{k=1}^{m}\left[\frac{1}{p}\sum_{j=1}^{p}(\lambda_{jk}^2)^2-\left\{\frac{1}{p}\sum_{j=1}^{p}\lambda_{jk}^2\right\}^2\right].$$

Maximizing this variance has the effect of driving the squared loadings to their extremes, ie. towards zero or one. Clearly other criteria can, and have, been devised which also push loadings towards zero or ± 1. Varimax is an orthogonal rotation method, but some of the other criteria which appear in standard packages allow oblique (non-orthogonal) factors.

References

1 Davenport Ellerby DR. *A statistical investigation of the treatment of certain brain damaged patients*. Unpublished MSc dissertation: University of Kent, Canterbury, 1980.

2 Bartlett JR, Neil-Dwyer G, Davenport-Ellerby DR, Morgan BJT. The rôle of the scanner in the care of the injured. Unpublished manuscript, 1981.

3 Jolliffe IT. *Principal component analysis*. New York: Springer-Verlag, 1986.

4 Diaconis P, Efron B. Computer-intensive methods in statistics. *Scientific American*, 1983; **248**: 96–108

5 Krzanowski WJ. *Principles of multivariate analysis, a user's perspective*. Oxford: Clarendon Press, 1988.

6 Gabriel KR, Odoroff CL. Biplots in biomedical research. *Stat Med* 1990; **9**: 469–85.

7 Lawley DN, Maxwell AE. *Factor analysis as a statistical method*, 2nd edition. London: Butterworth, 1971.

8 Hotelling H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 1933; **24**:417–41; 498–520.

9 Pearson K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 1901; 2: 559–72.

10 Spearman C. General intelligence, objectively determined and measured. *American Journal of Psychology*, 1904; **15**: 201–93.

11 Gifi A. *Nonlinear multivariate analysis*. Wiley: Chichester, 1990.

12 Everitt BS. *Statistical methods for medical investigations*. Bristol: Arnold, 1989.

13 Albert A, Harris EK. *Multivariate interpretation of clinical laboratory data*. New York: Dekker, 1987.

14 Ross CA, Joshi S, Currie R. Dissociative experiences in the general population: a factor analysis. *Hospital and Community Psychiatry* 1991; **42**: 297–301.

15 Henderson WG, Fisher SG, Cohen N, Waltzman S, Weber L, the VA co-operative study group on cochlear implantation. Use of principal components analysis to develop a composite score as a primary outcome variable in a clinical trial. *Controlled Clinical Trials* 1990; **11**: 199–214.

16 Yamamura K, Onodera S, Sasaki N, Matsumoto H, Nakano H, Makino M. Principal component analysis of various respiratory function tests: the relationship between factor score and severity of pulmonary circulatory disorder in chronic obstructive pulmonary disease. *Respiratory Medicine* 1991; **85**: 34–43.

17 Duflou H, Maenhaut W, DeReuck J. Application of principal component and cluster analysis to the study of the distribution of minor and trace elements in normal human brain. *Chemometrics and Intelligent Laboratory Systems* 1990; **9**: 273–86.

18 Andrews DF, Herzberg AM. *Data: a collection of problems from many fields for the student and research worker*. New York: Springer-Verlag, 1985.

19 Krzanowski WJ. Between-groups comparison of principal components. *Journal of the American Statistical Association* 1979; **74**: 703–707 (correction in **76**:1022).

20 Jeffers JNR. Two case studies in the application of principal component analysis. *Applied Statistics* 1967; **16**: 225–36.

21 Moser CA, Scott W. *British Towns*. Edinburgh: Oliver and Boyd, 1961.

22 Jolliffe IT, Jones B, Morgan BJT. Utilising clusters: a case study involving the elderly. *Journal of the Royal Statistical Society, Series A* 1982; **145**: 224–36.

23 Jolliffe IT. Discarding variables in a principal component analysis, I: Artificial data. *Applied Statistics* 1972; **21**: 160–73.

24 Gower JC. Some distance properties of latent roots and vector methods used in multivariate analysis. *Biometrika* 1966; **53**: 325–38.

25 Beumont PJV, Booth AL, Abraham SF, Griffiths DA, Turner TR. A temporal sequence of symptoms in patients with Anorexia Nervosa. Unpublished manuscript, 1981.

26 Banfield CF, Gower JC. A note on the graphical representation of multivariate binary data. *Applied Statistics* 1980; **29**: 238–45.

27 Morgan BJT, Ray APG. Non-uniqueness and inversions in cluster analysis. Unpublished manuscript, 1990.

28 Digby PGN, Kempton RA. *Multivariate analysis of ecological communities*. London: Chapman and Hall, 1987.

29 Jolliffe IT. Rotation of ill-defined principal components. *Applied Statistics* 1989; **38**: 139–47.

30 Huber P. Projection pursuit (with discussion). *Annals of Statistics* 1985; **13**: 435–525.

31 Jones MC, Sibson R. What is projection pursuit? (with discussion). *Journal of the Royal Statistical Society, Series A* 1987; **150**: 1–36.

32 Friedman JH. Exploratory projection pursuit. *Journal of the American Statistical Association* 1987; **82**: 249-66.

33 Morton SC. Interpretable projection pursuit.

Technical Report 45106. Department of Statistics, Stanford University, 1989.

34 Pack P, Jolliffe IT, Morgan BJT. Influential observations in principal component analysis: a case study. *Journal of Applied Statistics* 1988; **15**: 39–52.

35 Critchley F. Influence in principal component analysis. *Biometrika* 1985; 72: 627–36.

36 Tanaka Y, Odaka Y. Influential observations in principal factor analysis. *Psychometrika* 1989; **54**: 475–85.

37 Flury B. *Common principal components and related multivariate models*. New York: Wiley, 1988.

38 Flury B, Neuenschwander B. Principal components and proportionality in patterned covariance matrices. Unpublished manuscript, 1991

39 Chatfield C, Collins AJ. *Introduction to multivariate analysis*. London: Chapman and Hall, 1980: 89.

40 Jolliffe IT. Discarding variables in a principal component analysis, II: real data. *Applied Statistics*, 1973: **22**: 21–31.

41 Marx BD, Smith EP. Principal component estimation for generalized linear regression. *Biometrika* 1990; **77**: 23–31.

42 Jolliffe IT, Jones B, Morgan BJT. Comparison of cluster analyses of the English Personal Social Services Authorities. *Journal of the Royal Statistical Society, Series A* 1986; **149**: 253–70.

43 Chang W-C. On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics* 1983; **32**: 267–75.

44 Morrison DF. *Multivariate statistical methods* 2nd edition. Tokyo: McGraw-Hill Kogakusha, 1976.

45 Cattell RB. *The scientific use of factor analysis in behavioral and life sciences*. New York: Plenum Press, 1978.

46 Richman MB. Rotation of principal components. *Journal of Climatology*, 1986; **6**: 293–335.