Loading and correlations in the interpretation of principle compenents

Article *in* Journal of Applied Statistics · January 1995

DOI: 10.1080/757584614

CITATIONS
161

READS
215

2 authors, including:

Ian T. Jolliffe
University of Exeter
149 PUBLICATIONS **36,872** CITATIONS

SEE PROFILE

Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cjas20

Loading and correlations in the interpretation of principle compenents

Jorge Cadima Departamento de Matematica [a] & Ian T. Jolliffe Department of Mathematical Sciences [b]
[a] Departamento de Matematica , Instituto Superior de Agronomia
[b] Department of Mathematical Sciences , University of Aberdeen
Published online: 05 Jun 2011.

To cite this article: Jorge Cadima Departamento de Matematica & Ian T. Jolliffe Department of Mathematical Sciences (1995) Loading and correlations in the interpretation of principle compenents, Journal of Applied Statistics, 22:2, 203-214, DOI: 10.1080/757584614

To link to this article: http://dx.doi.org/10.1080/757584614

Loadings and correlations in the interpretation of principal components

JORGE CADIMA[1] **& IAN T. JOLLIFFE**[2], [1]*Departamento de Matemática, Instituto Superior de Agronomia, Portugal and* [2]*Department of Mathematical Sciences, University of Aberdeen, UK*

SUMMARY *Principal components (PC) are often interpreted by looking at the loadings for each variable. Variables with a small-magnitude loading are ignored and the given PC is then approximated by the linear combination involving only the remaining variables. It is argued that this procedure is potentially misleading in various respects. Examples are given and alternatives are suggested.*

1 Introduction

The interpretation or 'reification' (Krzanowski, 1988) of principal components (PCs) is often carried out based on the notion that it is the p original variables, or some subset of them, which can provide meaning to a PC. Thus, attention is focused on the linear combinations of the original variables which define the PCs. A first stage towards interpretation assesses whether or not a given PC can be adequately approximated by fewer than p variables. Those terms in a linear combination which have small-magnitude coefficients (variable loadings) are ignored. A second stage then focuses on this 'truncated PC', which is viewed as either a weighted average or a contrast of those variables which were retained, depending on whether or not the loadings are all of the same sign. A 'meaningful' PC is a PC for which the variables retained can be viewed as having a common feature of relevance to the problem which is being studied.

Interpretation of PCs is somewhat subjective and several authors have expressed reservations about it (see, for example, Jolliffe, 1986; Kendall, 1980; Chatfield & Collins, 1980). These criticisms qualify rather than reject the practice of interpreting PCs. However, with few exceptions (see, for example, Section 7.3 in Jackson (1991)), they have not focused on what we have called the first stage of the process, which is much less trivial than is usually thought.

0266-4763/95/020203-12

We shall see in Section 4, using three examples, that simple 'truncation' can be misleading. For example, one of the data sets, with four variables, has a component which is $-0.153\mathbf{x}_1 + 0.251\mathbf{x}_2 - 0.487\mathbf{x}_3 + 0.822\mathbf{x}_4$. The simplistic approach would approximate this component by $-0.487\mathbf{x}_3 + 0.822\mathbf{x}_4$, or something simpler, such as $0.8\mathbf{x}_4 - 0.5\mathbf{x}_3$, but a much better approximation is unexpectedly given by $-0.575\mathbf{x}_1 + 0.608\mathbf{x}_4$. Before considering the examples, it is necessary to discuss some background to the problem. Although the core results in this paper are also relevant for a probabilistic (inferential statistics) setting, the discussion is framed in a linear algebraic (descriptive statistics) context, which allows for a more intuitive visualization.

Section 2 summarizes the descriptive approach to PC analysis (PCA), and some of the associated geometry. These concepts help in interpreting loadings and correlations, as described in Section 3. The idea of regressing PCs on variables is introduced along with the examples in Section 4, and some conclusions are drawn in Section 5.

2 A descriptive setting

The rows of any given $n \times p$ (individuals vs variables) data matrix can be viewed as points in the p-dimensional Euclidean space $\mathbb{R}^p$, where each axis corresponds to a variable. These n points (and the vectors which they define with the origin) represent the n individuals. An alternative way of viewing the $n \times p$ data matrix is as a scatter or configuration of p points in $\mathbb{R}^n$, with each axis associated with an individual and each of the p points or vectors representing a variable. This dual interpretation of an $n \times p$ data matrix has been highlighted by many authors, such as Gower (1966), Cailliez and Pagès (1976), Kendall (1980), Volle (1981), Lebart *et al.* (1982 and 1984), Ramsay (1982), Escoufier (1987) and Krzanowski (1988).

A key factor in the importance of the dual representation is the interplay of geometrical and statistical concepts in both spaces. Let us consider an $n \times p$ data matrix $\mathbf{Y}$, its column-centred counterpart $\mathbf{X}$ and the matrix $(1/n^{1/2})\mathbf{X}$. The usual inner product between any two columns of $(1/n^{1/2})\mathbf{X}$, i.e.

$$\left\langle \frac{1}{n^{1/2}}\mathbf{x}_i, \frac{1}{n^{1/2}}\mathbf{x}_j \right\rangle = \frac{1}{n}\mathbf{x}_i'\mathbf{x}_j$$

is the covariance between the original variables from which they were derived (the variables $\mathbf{y}_i$ and $\mathbf{y}_j$). The norm or length defined by that inner product, i.e. the l_2 norm

$$\left\| \frac{1}{n^{1/2}}\mathbf{x}_i \right\| = \left(\frac{1}{n}\langle \mathbf{x}_i, \mathbf{x}_j \rangle \right)^{1/2}$$

is the standard deviation of the corresponding variable ($\mathbf{y}_i$). Also, the cosine of the angle between any two columns, i.e.

$$\cos\left(\frac{1}{n^{1/2}}\mathbf{x}_i, \frac{1}{n^{1/2}}\mathbf{x}_j \right) = \frac{\langle \mathbf{x}_i, \mathbf{x}_j \rangle}{\|\mathbf{x}_i\| \, \|\mathbf{x}_j\|}$$

is the correlation between the respective variables $\mathbf{y}_i$ and $\mathbf{y}_j$.

In many multivariate statistics techniques, including PCA, linear combinations of variables play a key role. These new variables can be represented in $\mathbb{R}^n$ as the resultants of those linear combinations of the p vectors that represent the original variables. This will provide a notion of the new variable's standard deviation (from its length) and correlation (cosine of the angle formed) with any other variable. Therefore, geometric intuition becomes a valuable aid in understanding and working with such techniques.

3 Loadings and PCs

When 'near-zero' loadings are ignored for the purpose of interpretation, they are effectively being replaced by exactly zero loadings. The resulting 'truncated PC' is then a vector in the subspace of $\mathbb{R}^n$ spanned by the k retained variables. The rationale for ignoring 'near-zero' loadings is that these will correspond to only minor displacements in the direction of the variables which they multiply, so can be safely discarded. We will see in this section that this premise is not generally true for covariance matrix PCs, since it does not take into account the variances (i.e. the lengths in $\mathbb{R}^n$) of each variable. Another problem with the method, for both covariance and correlation matrix PCs, is that it does not take into account the pattern of correlations (i.e. the relative positions in $\mathbb{R}^n$) of the variables.

Example 1. Kendall (1980, p. 20) discussed a data set consisting of $n = 20$ samples of soil for which $p = 4$ measurements are taken: silt content ($\mathbf{y}_1$), clay content ($\mathbf{y}_2$), organic matter ($\mathbf{y}_3$) and acidity ($\mathbf{y}_4$). The scale of the fourth variable (pH) is not comparable with those of the other variables (percentages) but we follow Kendall's use of a covariance matrix PCA. The loadings for the centred variables $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$ on the four PCs are reproduced in Table 1. The fact that the loadings in the main diagonal of Table 1 are all very large is a well-known consequence of the fact that the variances of the four variables (75.75, 13.13, 0.61 and 0.08 respectively) are of very different sizes (see Jolliffe, 1986).

Kendall (1980) considered the first PC and stated, "From the size of the coefficients ... we might provisionally call it silt content" (p. 20). For the second component, Kendall suggested that "the coefficients would lead us to regard it as dominated by [$\mathbf{x}_2$], namely clay". Most authors would probably also focus on the very high loadings in the main diagonal of Table 1 and consider that each PC is associated with a single variable. This conclusion would follow, for example, from using the criterion of Jeffers (1967) of retaining, for each PC, those variables whose loadings exceeds 70% of the largest loading for that PC (regardless of signs).

TABLE 1. Eigenpairs of the soil data's covariance matrix

	Eigenvectors			
Variable	$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$	$\mathbf{a}_4$
$\mathbf{x}_1$	0.956	−0.288	−0.059	0.006
$\mathbf{x}_2$	0.294	0.945	0.142	−0.018
$\mathbf{x}_3$	0.015	−0.154	0.979	−0.136
$\mathbf{x}_4$	0.001	−0.002	0.137	0.991
Eigenvalue	82.308	6.739	0.448	0.246
Percentage accounted	91.72	7.51	0.50	0.27

The conclusion that the second PC in Kendall's soil data is 'dominated' by the clay content ($\mathbf{x}_2$) should mean that $\mathbf{x}_2$ and the second PC are very 'close', by some criterion. One possible criterion is to require that the PC and its approximation be strongly correlated, i.e. that the angle between the vectors representing them in $\mathbb{R}^n$ be small. Another criterion would be to require that the standard Euclidean distance between the PC and its approximation be small, where 'small' is judged relative to the size (standard deviation) of the PC. This second criterion compares the PC scores and the 'truncated scores', dividing the sum of squares of their differences by the sum of squares of the ('full') scores and then taking the square root of that ratio.

It is well known (Mardia *et al.*, 1979) that the correlation between the ith variable and the jth PC is given by

$$\rho_{ij} = a_{ij}\left(\frac{\lambda_j}{s_{ii}}\right)^{1/2} \tag{1}$$

where a_{ij} is the loading for the ith variable in the jth PC, λ_j is the eigenvalue associated with that PC and s_{ii} is the variance of the ith variable. In Kendall's soil data, the correlation between the second PC and the second variable is only 0.6672—far from the convincing performance which the very high loading (0.945) would suggest. The situation is even worse if measured in terms of the distance between the PC and the truncated PC, since this distance is almost as large as the PC itself (the ratio is 0.9766). The variable $\mathbf{x}_2$ actually has a higher correlation with the first PC (0.7353) than with the second PC (despite having a loading of only 0.294 for that first PC).

The correlations between each PC and all the variables are given in Table 2. These correlations suggest several points:

(1) similar loadings, even very large ones, may translate into very different correlations between those variables and the PCs (see variable $\mathbf{x}_1$ in the first, and variable $\mathbf{x}_2$ in the second PCs);
(2) very different loadings may be associated with similarly correlated variables and PCs (see $\mathbf{x}_2$ in the first and second PCs);
(3) a given PC may be similarly correlated with a high and with a low loading variable (see the second PC and variables $\mathbf{x}_2$ and $\mathbf{x}_3$);
(4) the ranking of the loadings need not correspond to the ranking of the variables' correlations with the PC (see the second PC);
(5) the interpretability of a PC does not depend on the percentage of the total variability it accounts for (0.27% in the case of the fourth PC, which is very strongly correlated with the fourth variable).

TABLE 2. Correlations between variables and PCs for Kendall's soil data

	PC1	PC2	PC3	PC4
$\mathbf{x}_1$	0.9963	−0.0859	−0.0045	0.0004
$\mathbf{x}_2$	0.7353	0.6772	0.0261	−0.0025
$\mathbf{x}_3$	0.1737	−0.5107	0.8376	−0.0862
$\mathbf{x}_4$	0.0239	−0.0101	0.1840	0.9826

The variances of the variables and components must always be taken into account when judging the loadings, as is highlighted by equation (1). The loadings in correlation matrix PCA are sometimes multiplied by $\lambda_j^{1/2}$, so effectively giving the correlation of each variable with the PC. No such simple rescaling is possible in covariance matrix PCA, where the loadings for any given PC would have to be multiplied by different scalars to obtain the correlations of equation (1).

What is happening is best visualized by considering the PCs as the linear combinations of vectors in $\mathbb{R}^n$. It is not just the coefficients of the linear combination but also the size (variance) and the relative positions of the vectors (correlations between the variables) which determine the results. In particular, the first PC of any covariance matrix PCA is indeed associated with the large-loading variables, but to an even larger degree than the loadings alone appear to indicate, since the largest coefficients of the first eigenvector of a covariance matrix correspond to the variables with largest variances. For the remaining PCs, no such general rule can be established. Also, for correlation matrix PCA, where all the vectors that represent variables are of equal length in $\mathbb{R}^n$, the relative positions of (correlations between) the variables will continue to blur the picture provided by loadings.

However, the above discussion has only addressed part of the problem, since correlations between individual variables and any given PC are not sufficient to assess the validity of approximations based on a choice of $k > 1$ variables. In general, it is not possible to determine how close a PC is to its 'truncated' approximations merely by looking at the correlations between individual variables and PCs.

Let us consider the third PC in the same example. Judging from loadings alone, this third PC appears to be dominated by organic matter ($\mathbf{x}_3$). However, the correlation between these two variables (0.8376), although much better than that between $\mathbf{x}_2$ and the second PC, may still be considered to be unsatisfactory if we are to speak of a PC 'dominated' by $\mathbf{x}_3$ (the angle between them in $\mathbb{R}^n$ is approximately 33°). The distance from the PC to the 'truncated PC', i.e. $0.979\mathbf{x}_3$, is actually larger than the PC itself (their ratio is 1.2439). If we are to add new terms to the 'truncated PC' according to their loadings, then the most sensible choice appears to be the terms in both $\mathbf{x}_2$ and $\mathbf{x}_4$, whose loadings are nearly equal. If we are to add the term in the variable second best correlated with the PC, then $\mathbf{x}_4$ alone seems adequate (its correlation with the PC is 0.1840, whereas that of $\mathbf{x}_2$ is 0.0261 and that of $\mathbf{x}_1$ is -0.0045). In both cases, the triplet $\{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$ seems to be the best three-variable subset for approximating the third PC. However, the correlation between $v = 0.142\mathbf{x}_2 + 0.979\mathbf{x}_3 + 0.137\mathbf{x}_4$ and the third PC actually drops to 0.7923—a worse approximation than with $\mathbf{x}_3$ alone. In contrast, if the term in $\mathbf{x}_4$ is replaced with the term in $\mathbf{x}_1$ (giving the approximate PC $v = -0.059\mathbf{x}_1 + 0.142\mathbf{x}_2 + 0.979\mathbf{x}_3$), then the correlation with the PC grows to a substantial 0.9948 (an angle of just under 6° in $\mathbb{R}^n$). Distance-wise, this last approximation is also the best three-variable truncated PC. The implications is that $\mathbf{x}_1$ plays a significant role in ensuring a good approximation of the third PC, despite its negligible loading (-0.059) and correlation with that PC (-0.0045). The size (standard deviation) and position (correlation) of this first variable relative to other variables are being taken into account.

The soil data's third PC illustrates that adding new terms to a truncated PC might actually make the approximation worse. To understand this apparently strange conclusion, let us generalize equation (1) to give the correlation between any PC, giving as $\mathbf{X}\mathbf{a}_j$, and a truncated PC given by $\mathbf{X}_\kappa \mathbf{a}_j^\kappa$, where $\mathbf{X}_\kappa$ is the sub-

matrix of $\mathbf{X}$ obtained by retaining only k of its columns (the column numbers forming the set of integers κ) and $\mathbf{a}_j^\kappa$ is the subvector of $\mathbf{a}_j$ which results from retaining only the k coefficients associated with the variables which were retained. We have (see Appendix A)

$$r_t = \text{corr}(\mathbf{X}\mathbf{a}_j, \mathbf{X}_\kappa \mathbf{a}_j^\kappa) = \frac{\lambda_j^{1/2} \mathbf{a}_j^{\kappa'} \mathbf{a}_j^\kappa}{(\mathbf{a}_j^{\kappa'} \mathbf{S}_\kappa \mathbf{a}_j^\kappa)^{1/2}} \tag{2}$$

where $\mathbf{S}_\kappa$ is the submatrix of the covariance matrix involving only the rows and columns corresponding to the retained variables (i.e. the covariance matrix determined by $\mathbf{X}_\kappa$). Discarding the smallest magnitude loadings, as is usually done, will maximize the numerator (for any fixed number k of retained variables). However, this advantage must be weighted against the consequences in the denominator, which are far more complex.

Considering the distance between the PC and the truncated PC, relative to the size (standard deviation) of the PC itself, we have (see Appendix A)

$$d_t = \frac{\|\mathbf{X}\mathbf{a}_j - \mathbf{X}_\kappa \mathbf{a}_j^\kappa\|}{\|\mathbf{X}\mathbf{a}_j\|} = \left(1 - 2\mathbf{a}_j^{\kappa'} \mathbf{a}_j^\kappa + \frac{\mathbf{a}_j^{\kappa'} \mathbf{S}_\kappa \mathbf{a}_j^\kappa}{\lambda_j}\right)^{1/2} \tag{3}$$

The last term in the parentheses shows that, for this criterion also, the best k-variable subset is not guaranteed just by discarding the $p-k$ variables with smallest loadings.

To summarize, loadings are not reliable for determining whether or not some subset of the p original variables can provide an acceptable truncated PC, particularly for covariance matrix PCA. Likewise, correlations between individual variables and PCs are not appropriate, except when judging the adequacy of single-variable approximations.

4 Regressing PCs on variables

Given a subset of k variables, the PC is then usually approximated by the truncated PC, i.e. by the sum of those terms of

$$\mathbf{X}\mathbf{a}_j = \sum_{i=1}^{p} a_{ij} \mathbf{x}_i$$

corresponding to the k selected variables. However, this choice does not generally give the best approximate PC using those k variables.

Approximating a given variable (the PC) with a linear combination of some set of k other variables is the problem which multiple linear regression addresses. Given the subset of k variables, the multiple linear regression equation provides the best approximation in both the least-squares sense of minimizing the distance to the PC and in the sense of maximizing the correlation with the PC. Thus, the truncated PCs with which we have worked in the previous section are suboptimal for both criteria. In our linear algebraic context, regression is the orthogonal projection of the PC on the subspace of $\mathbb{R}^n$ spanned by the k variables, and the correlation between the PC and its projection is the multiple correlation coefficient of the k variables with the PC (Draper & Smith, 1981).

In Appendix A, it is shown that the projection of PC ($\mathbf{X}\mathbf{a}_j$) onto the subspace spanned by the k variables whose indexes form the set κ, i.e. $\{\mathbf{x}_i\}_{i \in \kappa}$, is the vector

$\mathbf{X}_\kappa(\lambda_j \mathbf{S}_\kappa^{-1}\mathbf{a}_j^\kappa)$. Its correlation with the PC itself, i.e., the multiple correlation coefficient between the jth PC and the k variables defined by subset κ, is (see Appendix A)

$$r_m = \lambda_j^{1/2}(\mathbf{a}_j^{\kappa'}\mathbf{S}_\kappa^{-1}\mathbf{a}_j^\kappa)^{1/2} \tag{4}$$

The ratio r_t/r_m is the correlation between the truncated PC and the projected PC (see Appendix A). Furthermore, the distance from the PC to the projected PC, relative to the size of the PC, is merely the sine of the angle beween those two vectors in $\mathbb{R}^n$, i.e.

$$d_m = (1 - r_m^2)^{1/2} \tag{5}$$

As with equations (2) and (3), it is not necessary to know the data set to compute these values. All that is needed is the covariance matrix.

It should be noted that the truncated PC and the projected PC coincide (apart from a scaling factor) if and only if $r_t/r_m = 1$. An equivalent condition is that $\mathbf{a}_j^\kappa$ is an eigenvector of $\mathbf{S}_\kappa$ (see Appendix A). Thus, the truncated PC will only be as well correlated as the projected PC if the subvector $\mathbf{a}_j^\kappa$ of $\mathbf{S}$'s eigenvector $\mathbf{a}_j$ is also an eigenvector for the submatrix $\mathbf{S}_\kappa$.

It should also be noted that the multiple correlation r_m conveys all the relevant information concerning the projected PCs' behaviour in terms of both the correlation and the distance criteria. The situation with truncated PCs is not as simple, since both r_t and the magnitude of the loadings retained will affect the distance d_t. However, experience suggests that the values of d_t are often considerably larger than those of d_m. They can even be greater than unity (unlike d_m) if $r_t^2 < \mathbf{a}_j^{\kappa'}\mathbf{a}_j^\kappa/2$.

The multiple correlation coefficient should be used to decide which subset of the k variables can ensure a good approximate PC. Among its advantages are the following: (1) its simple geometric interpretation; (2) its previous use in choosing a subset of variables to provide a 'good' approximation for another variable (see, for example, Mardia *et al.*, 1979); (3) unlike with r_t, adding any new variables to the set of regressor variables can never decrease the value of the muliple correlation. However, the overriding consideration that makes the use of multiple correlations more appropriate is that this technique is associated with a method of approximation that provides an optimum under two different criteria, and it is not always the case that projected PCs and truncated PCs are similar, as will now be shown.

Example 2. A second data set—this time with a larger number of variables—was taken from Lebart *et al.* (1982, p. 283). It summarizes a study of yearly expenditures on foodstuffs by French families. There are $p = 7$ variables, corresponding to groups of foodstuffs: bread, vegetables, fruit, red meat, poultry, milk and wine. In reality, the $n = 12$ 'individuals' are the average expenditures of the sampled families for each of 12 categories: the combinations of each of three social groups (manual workers, non-manual workers, technical and managerial staff) with each of four family sizes (parents with 2, 3, 4 or 5 children). We shall not be unduly concerned with the somewhat unconventional nature of the 'individuals', because it is the numbers—and not their precise nature—which are of interest to us.

Although Lebart *et al.* chose to analyze the correlation matrix for this data set, a covariance matrix PCA is also appropriate. All the variables are in the same units (French francs) and separate rescalings of each variable do not make much sense. Table 3 gives the loadings for the first three (covariance matrix) PCs of this data

TABLE 3. Loadings for the first three PCs of the foodstuffs expenditures data

	Loading		
Variable	PC1	PC2	PC3
$\mathbf{x}_1$	0.073	0.576	0.404
$\mathbf{x}_2$	0.328	0.409	−0.292
$\mathbf{x}_3$	0.303	−0.100	−0.340
$\mathbf{x}_4$	0.753	−0.108	0.068
$\mathbf{x}_5$	0.465	−0.244	0.381
$\mathbf{x}_6$	0.091	0.632	−0.225
$\mathbf{x}_7$	−0.059	0.144	0.660
Eigenvalue	251928	24215	5733
Percentage accounted	88.00	8.46	2.00

set, as well as the eigenvalues and percentage variance accounted for by each of these PCs.

For each of the first three PCs, which account for over 98% of the total variance, those variables whose loading exceeds 0.30 in magnitude (the precise threshold is not crucial for the discussion that follows) will be retained. The choices that result are given in Table 4, together with the values of r_t, r_m, d_t and d_m for each case.

The first PC is very well approximated by the subset of high loading variables, with little difference between the truncated and the projected PCs. This appears to be largely due to the fact that both subsets include variable $\mathbf{x}_4$ which, by itself, is already an excellent 'approximate PC', forming an angle of under 4° with the PC in $\mathbb{R}^n$ (the correlation is 0.9977).

However, the results for the second and third PCs are much more interesting. Reasonably good approximations of both PCs can be obtained using the subsets selected, since the lowest of the multiple correlations (that of $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_6\}$ with the second PC) is 0.9641. However, in both cases, the approximations provided by the truncated PCs are significantly worse than those given by regression.

The implications of this are considerable. Not only might we reject some potentially adequate subsets of variables, on the basis of the poor performance of their truncated PCs, but a more serious problem arises: that of differing interpretations for a given PC, based on which method of approximation is chosen. Let us consider the approximation to the second PC provided by taking only those terms of $\mathbf{Xa}_2$ in the variables $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_6$, i.e. $\mathbf{v}_1 = 0.576\mathbf{x}_1 + 0.409\mathbf{x}_2 + 0.632\mathbf{x}_6$. This second PC might then be viewed as a weighted average of bread, vegetables and milk. However, a linear multiple regression of the PC on the same set of variables would produce the approximate vector which, after rescaling the coefficients to a comparable size (the same sum of squares as in $\mathbf{v}_1$), is given by $\mathbf{v}_2 = 0.546\mathbf{x}_1 - 0.266\mathbf{x}_2 + 0.727\mathbf{x}_6$. This approximation will lead us to view the PC as a contrast between bread and milk, on the one hand, and vegetables, on the other hand. Also $\mathbf{v}_2$'s correlation with the PC is significantly better (0.9641 as compared with 0.7656), so it is a more faithful reflection (scaling factors aside) of that PC.

Looking briefly at the third PC, the truncated PC using only $\mathbf{x}_1$, $\mathbf{x}_3$, $\mathbf{x}_5$ and $\mathbf{x}_7$ can be seen from Table 3 to contrast $\mathbf{x}_3$ (fruit) with the other three variables. However, the corresponding projected PC, whose correlation with the PC grows

TABLE 4. Indicators of quality of approximation of the first three PCs using projected and truncated PCs, based on subsets of high loading variables in the Lebart data set

PC	Variables in subset	r_t	r_m	d_t	d_m
1	$\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5$	0.9996	1.0000	0.0338	0.0061
2	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_6$	0.7656	0.9641	0.7615	0.2657
3	$\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_5, \mathbf{x}_7$	0.7725	0.9790	0.7203	0.2036

from 0.7725 to 0.9790, is $\mathbf{v} = -0.057\mathbf{x}_1 - 0.481\mathbf{x}_3 + 0.408\mathbf{x}_5 + 0.675\mathbf{x}_7$. Thus, bread ($\mathbf{x}_1$) has 'changed sides' in the contrast, although with a relatively small coefficient. This small coefficient, coupled with $\mathbf{x}_1$'s unimpressive variance (10 524 in a covariance matrix whose trace is 286 272), suggests that the third, fifth and seventh variables are perhaps sufficient for a good approximation. In fact, the multiple correlation of the third PC with $\mathbf{x}_3$, $\mathbf{x}_5$ and $\mathbf{x}_7$ turns out to be only marginally smaller, i.e. 0.9752. Incidentally, this subset leaves out the variable with the second largest magnitude loading ($\mathbf{x}_1$).

Example 3. Unlike previous examples, we analyze here a correlation matrix PCA and show that problems can arise even when all the vectors that represent variables are of equal length, if a loadings-only approach to the interpretation of PCs is used. Arnold and Collins (1993) gave the ratings by an 'Assessor A' of eight commercial port wines on four attributes.

The third correlation matrix PC for this data set accounts for 3.4% of the total variability. Its loadings are -0.153, 0.251, -0.487 and 0.822. The correlation of this PC with the fourth variable alone is low (0.3049), despite the significant loading. If we also consider the third variable, which has the second highest magnitude loading, then the correlation between the resulting truncated PC ($\mathbf{v} = -0.487\mathbf{x}_3 + 0.822\mathbf{x}_4$) and the PC is still only 0.6648. A better approximation to the PC can be obtained using the same two-variable subset, since the projected PC ($\mathbf{v} = -0.658\mathbf{x}_3 + 0.693\mathbf{x}_4$, with the same sum of squared coefficients as above) has a correlation of 0.8039 with the PC. However, the highest multiple correlation of the PC with any two-variable subset occurs when we select the first and fourth variables. In fact, the correlation between the PC and the projected PC $\mathbf{v} = -0.575\mathbf{x}_1 + 0.608\mathbf{x}_4$ is 0.9165. Again, we have ended up with a subset which includes the variable with the smallest magnitude loading.

The general conclusion is that the use of the multiple correlation coefficient in assessing the performance of a subset of variables with regards to a given PC is not just a question of minor improvements, and 'truncating' a PC is a potentially misleading means of assessing how a given subset of variables combines for a best approximation of a PC. The problem is all the more serious if a fairly high correlation between the PC and its 'truncation' is not achieved with a particular choice of variables. Interpretation of PCs based on such truncated PCs may then be flawed.

5 Discussion and conclusions

We have attempted to show that more rigour and quantification are called for when PCs are associated with subsets of variables, as a prelude to interpretation. The traditional practice of selecting such subsets from their corresponding loadings

and of truncating PCs as a means to obtain approximate PCs may be seriously unreliable. One of the reasons for this is that it is not just loadings but also the size (standard deviation) of each variable which determines the importance of that variable in the linear combination. This suggests that we could rewrite the PC as a linear combination of vectors of equal size but 'correcting' the loadings in this fashion does not solve all our problems. In fact, the relationship between the variables (expressed through the pattern of correlations which expresses the relative position of each variable in $\mathbb{R}^n$) also affect the result. Multiple correlation coefficients and the regression of PCs on subsets of the variables are more appropriate alternatives.

Regressing PCs on variables—which is a curious inversion of the relationship which originally motivated Hotelling's approach to PCA (Hotelling, 1933)—raises numerous technical difficulties, particularly when the number of variables is not very small. Previous work on multiple linear regression provides useful background material to tackle such difficulties.

On a more general level, the discussion stresses the importance of clearly distinguishing between PCs and their vectors of loadings. The two sets of vectors are often confused—to the point of using the term 'principal components' in both cases—and, although there is a very strong relationship between them, obscuring their distinct nature can be misleading.

Finally, the choice of subsets of variables may also be affected by a different (if related) purpose: that of providing a good approximation to 'principal subspaces' (i.e. subspaces spanned by various PCs) rather than to individual PCs. This topic, which can be dealt with in a similar spirit to that used here, will be discussed in a separate paper.

Acknowledgements

The work of one of us (JC) was supported in part by the Calouste Gulbenkian Foundation and by the European Community's SCIENCE programme.

Correspondence: Mr J. Cadima, Departamento de Matematica, Instituto Superior de Agronomia, Tapada da Ajuda, 1399 Lisboa Codex, Portugal.

REFERENCES

Arnold, G. M. & Collins, A. J. (1993) Interpretation of transformed axes in multivariate analysis, *Applied Statistics*, 42, pp. 381–400.

Cailliez, F. & Pages, J.-P. (1976) *Introduction à l'analyse des données* (Paris, Société de Mathématiques Appliquées et de Sciences Humaines).

Chatfield, C. & Collins, A. J. (1980) *Introduction to Multivariate Analysis* (London, Chapman and Hall).

Draper, N. R. & Smith, H. (1981) *Applied Regression Analysis*, 2nd edn (New York, Wiley).

Escoufier, Y. (1987) The duality diagram: a means for better practical applications. In: P. Legendre & L. Legendre (Eds), *Developments in Numerical Ecology* (Berlin, Springer).

Gower, J. C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis, *Biometrika*, 53, pp. 325–338.

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

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

Jeffers, J. N. R. (1967) Two case studies in the application of principal component analysis, *Applied Statistics*, 16, pp. 225–236.

Jolliffe, I. T. (1986) *Principal Component Analysis* (New York, Springer).

KENDALL, M. (1980) *Multivariate Analysis*, 2nd edn (London, Charles Griffin).
KRZANOWSKI, W. (1988) *Principles of Multivariate Analysis: A User's Perspective* (Oxford, Clarendon Press).
LEBART, L., MORINEAU, A. & FÉNELON, J.-P. (1982) *Traitement des données statistiques* (Paris, Dunod).
LEBART, L., MORINEAU, A. & WARWICK, K. M. (1984) *Multivariate Descriptive Statistical Analysis* (New York, Wiley).
MARDIA, K. V., KENT, J. T. & BIBBY, J. M. (1979) *Multivariate Analysis* (London, Academic Press).
RAMSAY, J. O. (1982) When the data are functions, *Psychometrika*, 47, pp. 379–396.
VOLLE, M. (1981) *Analyse des données*, 2nd edn (Paris, Economica).

Appendix A

The truncated PC can be written as $\mathbf{X}_\kappa \mathbf{a}_j^\kappa$, where $\mathbf{X}_\kappa$ is the $n \times k$ submatrix of $\mathbf{X}$ obtained by deleting the columns not associated with variables in the set κ. The projected PC is given by $[\mathbf{X}_\kappa(\mathbf{X}_\kappa'\mathbf{X}_\kappa)^{-1}\mathbf{X}_\kappa']\mathbf{X}\mathbf{a}_j$. Denoting by $\mathbf{I}_\kappa$ the $p \times k$ submatrix of the $p \times p$ identity obtained by retaining only those columns whose rank numbers are in κ, we see that $\mathbf{X}_\kappa = \mathbf{X}\mathbf{I}_\kappa$. Hence, the projected PC becomes

$$\mathbf{X}_\kappa(\mathbf{X}_\kappa'\mathbf{X}_\kappa)^{-1}\mathbf{I}_\kappa'\mathbf{X}'\mathbf{X}\mathbf{a}_j = \mathbf{X}_\kappa\mathbf{S}_\kappa^{-1}\mathbf{I}_\kappa'\mathbf{S}\mathbf{a}_j = \lambda_j\mathbf{X}_\kappa\mathbf{S}_\kappa^{-1}\mathbf{I}_\kappa'\mathbf{a}_j = \lambda_j\mathbf{X}_\kappa\mathbf{S}_\kappa^{-1}\mathbf{a}_j^\kappa$$

The correlation between the PC and the truncated PC is

$$r_t = \cos(\mathbf{X}\mathbf{a}_j, \mathbf{X}_\kappa\mathbf{a}_j^\kappa) = \frac{\langle \mathbf{X}\mathbf{a}_j, \mathbf{X}_\kappa\mathbf{a}_j^\kappa\rangle}{\|\mathbf{X}\mathbf{a}_j\|\,\|\mathbf{X}_\kappa\mathbf{a}_j^\kappa\|} = \frac{\mathbf{a}_j'\mathbf{S}\mathbf{I}_\kappa\mathbf{a}_j^\kappa}{\lambda_j^{1/2}(\mathbf{a}_j^{\kappa\prime}\mathbf{S}_\kappa\mathbf{a}_j^\kappa)^{1/2}} = \frac{\lambda_j^{1/2}\mathbf{a}_j^{\kappa\prime}\mathbf{a}_j^\kappa}{(\mathbf{a}_j^{\kappa\prime}\mathbf{S}_\kappa\mathbf{a}_j^\kappa)^{1/2}}$$

The correlation between the PC and the projected PC is

$$r_m = \cos(\mathbf{X}\mathbf{a}_j, \lambda_j\mathbf{X}_\kappa\mathbf{S}_\kappa^{-1}\mathbf{a}_j^\kappa) = \frac{\mathbf{a}_j'\mathbf{S}\mathbf{I}_\kappa\mathbf{S}_\kappa^{-1}\mathbf{a}_j^\kappa}{\lambda_j^{1/2}(\mathbf{a}_j^{\kappa\prime}\mathbf{S}_\kappa^{-1}\mathbf{S}_\kappa\mathbf{S}_\kappa^{-1}\mathbf{a}_j^\kappa)^{1/2}} = \lambda_j^{1/2}(\mathbf{a}_j^{\kappa\prime}\mathbf{S}_\kappa^{-1}\mathbf{a}_j^\kappa)^{1/2}$$

The correlation between the truncated PC and the projected PC is

$$\cos(\mathbf{X}_\kappa\mathbf{a}_j^\kappa, \lambda_j\mathbf{X}_\kappa\mathbf{S}_\kappa^{-1}\mathbf{a}_j^\kappa) = \frac{\langle \mathbf{X}_\kappa\mathbf{a}_j^\kappa, \lambda_j\mathbf{X}_\kappa\mathbf{S}_\kappa^{-1}\mathbf{a}_k^\kappa\rangle}{\|\mathbf{X}_\kappa\mathbf{a}_j^\kappa\|\,\|\lambda_j\mathbf{X}_\kappa\mathbf{S}_\kappa^{-1}\mathbf{a}_j^\kappa\|} = \frac{\mathbf{a}_j^{\kappa\prime}\mathbf{a}_j^\kappa}{(\mathbf{a}_j^{\kappa\prime}\mathbf{S}_\kappa\mathbf{a}_j^\kappa)^{1/2}(\mathbf{a}_j^{\kappa\prime}\mathbf{S}_\kappa^{-1}\mathbf{a}_j^\kappa)^{1/2}} = \frac{r_t}{r_m}$$

For the projected PC to coincide (a multiplicative scaling factor aside) with its corresponding truncated PC, we must have, for some constant η, that

$$\mathbf{X}_\kappa\mathbf{a}_j^\kappa = \eta\mathbf{X}_\kappa\mathbf{S}_\kappa^{-1}\mathbf{a}_j^\kappa \Rightarrow \mathbf{X}_\kappa'\mathbf{X}_\kappa\mathbf{a}_j^\kappa = \eta\mathbf{S}_\kappa\mathbf{S}_\kappa^{-1}\mathbf{a}_j^\kappa \Leftrightarrow \mathbf{S}_\kappa\mathbf{a}_j^\kappa = \eta\mathbf{a}_j^\kappa$$

so that $\mathbf{a}_j^\kappa$ is an eigenvector of $\mathbf{S}_\kappa$. In contrast, if $\mathbf{a}_j^\kappa$ is an eigenvector of $\mathbf{S}_\kappa$, then it is an eigenvector of $\mathbf{S}_\kappa^{-1}$, so $\mathbf{X}_\kappa\mathbf{S}_\kappa^{-1}\mathbf{a}_j^\kappa$ can be written as $\eta\mathbf{X}_\kappa\mathbf{a}_j^\kappa$ for some scalar η.

The distance between the PC and the truncated PC, relative to the size of the PC, is

$$d_t = \frac{\|\mathbf{X}\mathbf{a}_j - \mathbf{X}_\kappa\mathbf{a}_j^\kappa\|}{\|\mathbf{X}\mathbf{a}_j\|} = \left(\frac{\lambda_j - 2\mathbf{a}_j'\mathbf{S}\mathbf{I}_\kappa\mathbf{a}_j^\kappa + \mathbf{a}_j^{\kappa\prime}\mathbf{S}_\kappa\mathbf{a}_j^\kappa}{\lambda_j}\right)^{1/2} = \left(1 - 2\mathbf{a}_j^{\kappa\prime}\mathbf{a}_j^\kappa + \frac{\mathbf{a}_j^{\kappa\prime}\mathbf{S}_\kappa\mathbf{a}_j^\kappa}{\lambda_j}\right)^{1/2}$$

The distance between the PC and the projected PC, relative to the size of the PC, is

$$d_{\mathrm{m}} = \frac{\|\mathbf{X}\mathbf{a}_j - \lambda_j \mathbf{X}_\kappa \mathbf{S}_\kappa^{-1} \mathbf{a}_j^\kappa\|}{\|\mathbf{X}\mathbf{a}_j\|} = \left(\frac{\lambda_j - 2\lambda_j \mathbf{a}_j^{\kappa'} \mathbf{S}\mathbf{I}_\kappa \mathbf{S}_\kappa^{-1} \mathbf{a}_j^\kappa + \lambda_j^2 \mathbf{a}_j^{\kappa'} \mathbf{S}_k^{-1} \mathbf{a}_j^\kappa}{\lambda_j} \right)^{1/2}$$

$$= (1 - \lambda_j \mathbf{a}_j^{\kappa'} \mathbf{S}_\kappa^{-1} \mathbf{a}_j^\kappa)^{1/2} = (1 - r_{\mathrm{m}}^2)^{1/2} = \sin(\mathbf{X}\mathbf{a}_j, \lambda_j \mathbf{X}_\kappa \mathbf{S}_\kappa^{-1} \mathbf{a}_j^\kappa)$$