On the interpretation of the smallest principal component

R. A. REYMENT

Reyment, R. A. 1979 4 30: On the interpretation of the smallest principal component. *Bulletin of the Geological Institutions of the University of Uppsala,* N.S., Vol. 8, pp. 1—4. Uppsala. ISSN 0302-2749.

The eigenvector attached to the smallest eigenvalue (i.e. the principal component with the least variance), in cases where this differs but insignificantly from nought, presents a linear combination of variables which is invariant in the sample. Some examples of the interpretation of such small principal components for planktonic and benthic foraminifers, living grasshoppers and mineral chemistry are given. For the biological examples, the smallest principal component can provide useful information on the occurrence of invariant growth relationships. In mineral chemistry, proportional relations in elements show up in the least principal component.

R. A. Reyment, Paleontologiska Institutionen, Uppsala Universitet, Box 558, S-751 22, Uppsala, Sweden, 14th December, 1978.

Introduction

Ever since Hotelling (1933) introduced the use of the eigenvalues and eigenvectors of a covariance matrix into the statistical literature under the name of *principal component analysis*, interest seems always to have been directed towards interpreting the first few eigenvectors, connected to most of the variance. The reason for doing this is obvious, for the investigator will normally want to learn as much as possible about those linear combinations of his variables that are providing most of the variability in his material.

There is, however, another kind of question that ought to be of interest, but which seems to have remained unasked, although some mathematical statisticians have wondered briefly over this (Gnanadesikan and Wilk 1969; Gower 1967). This question concerns the possibility offered by the eigenvector attached to the smallest (zero or almost zero) eigenvalue of finding a linear combination of variables which is invariant in the material. That is, that combination which is constant, or almost constant, for variables measured in the same metric.

The justification for this idea may perhaps not be immediately obvious. Gnanadesikan and Wilk (1969), in a geometrically constructed example, showed the way in which the smallest eigenvalue and its vector can be employed for determining a structural relationship.

The purpose of the present paper is to point out the possibilities seemingly offered by the smallest principal component in geology. It is not intended to be an exhaustive final word on the subject for, in order to be able to access all the aspects involved, a large-scale programme of simulation studies would be necessary.

This note is a slightly expanded version of a paper published in Russian in a special volume of the Soviet Academy of Sciences, issued in honour of Professor A. B. Vistelius on the occasion of his sixtieth birthday (Reyment, 1978).

Component Analysis

Following Jöreskog, Klovan and Reyment (1976), we shall consider what is meant by component analysis, confining ourselves to a development of the fixed case model.

Consider a data matrix, $\mathbf{Y}$, expressed in deviate form,

$$\mathbf{Y} = \mathbf{F}\mathbf{A}' + \mathbf{E}, \qquad (1)$$

where $\mathbf{F}$ is a matrix of factor scores, $\mathbf{A}$ the matrix of factor loadings and $\mathbf{E}$ is a matrix of residuals.

If we assume that all elements of $\mathbf{Y}$ have been divided by $\sqrt{N}$ the covariance matrix, $\mathbf{S}$, is

$$\mathbf{S} = \mathbf{Y}'\mathbf{Y}.$$

This model may be fitted by applying the method of least squares to the data matrix, $\mathbf{Y}$. It is therefore necessary to determine the matrices $\mathbf{F}_{(N\times k)}$ and $\mathbf{A}_{(p\times k)}$ for a given $k < p$, such that the sum of squares of all the elements of the matrix

$$\mathbf{E} = \mathbf{Y} - \mathbf{F}\mathbf{A}' \qquad (2)$$

is as small as possible. The solution to this problem is obtained as

$$\mathbf{F}\mathbf{A}' = g_1\mathbf{v}_1\mathbf{u}_1' + g_2\mathbf{v}_2\mathbf{u}_2' + \ldots + g_k\mathbf{v}_k\mathbf{u}_k',$$

being k terms of the singular value decomposition of **Y** corresponding to the k largest singular values $g_1, g_2, \ldots, g_k$. Writing

$$\mathbf{V}_k = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k],$$
$$\mathbf{U}_k = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k],$$

and,

$$\mathbf{G}_k = \text{diag}(g_1, g_2, \ldots, g_k),$$

the solution is,

$$\mathbf{FA}' = \mathbf{V}_k\mathbf{G}_k\mathbf{U}_k' .$$

This provides the product $\mathbf{FA}'$, but it does not give a unique solution for **F** and **A** themselves.
Two different solutions are available for **A** and **F**, both of which will be presented here, but only one of which is useful for treating our specific problem. It should be noted, however, that these solutions are essentially the same, differing only in the way which the factors are scaled.

The first solution is

$$\mathbf{F} = \mathbf{V}_k,\ \mathbf{A} = \mathbf{U}_k\mathbf{G}_k. \qquad (3)$$

Inasmuch as the column sums of matrix **Y** are nought, so are those of $\mathbf{V}_k$. Therefore, **F** will be in deviate form and the covariance matrix of the factors, in the sample, will be

$$\mathbf{F}'\mathbf{F} = \mathbf{V}_k'\,\mathbf{V}_k = \mathbf{I},$$

which indicates that the factors are uncorrelated and standardized. For the factor loadings matrix we have,

$$\mathbf{A}'\mathbf{A} = \mathbf{G}_k\mathbf{U}_k'\,\mathbf{U}_k\mathbf{G}_k = \mathbf{G}_k^2 = \Lambda_k,$$

where Λ_k is a diagonal matrix, of order $k \times k$, the diagonal elements of which are $\lambda_1 = g_1^2$, $\lambda_2 = g_2^2, \ldots, \lambda_k = g_k^2$, the eigenvalues of **S**, the covariance matrix. If matrix **E** in (1) is small, we have, approximately, that

$$\mathbf{Y} \approx \mathbf{FA}',$$

and covariance matrix **S** is approximately

$$\mathbf{S} = \mathbf{Y}'\mathbf{Y} \approx \mathbf{AF}'\mathbf{FA}' = \mathbf{AA}'.$$

Assuming that $N \geqq p$, one computes $\mathbf{A} = \mathbf{U}_k\,\Lambda_k^{1/2}$, which amounts to scaling each eigenvector so that the square of its length equals the corresponding eigenvalue. The estimate of **F** is then $\mathbf{YA}\Lambda_k^{-1}$.

For the second solution,

$$\mathbf{A} = \mathbf{U}_k,\ \mathbf{F} = \mathbf{V}_k\mathbf{G}_k. \qquad (4)$$

With this solution, the covariance matrix of the factors is

$$\mathbf{F}'\mathbf{F} = \mathbf{G}_k\mathbf{V}_k'\,\mathbf{V}_k\mathbf{G}_k = \mathbf{G}_k^2 = \Lambda_k,$$

which is diagonal, but with diagonal elements equal to the eigenvalues of **S** in descending order of magnitude. Hence, the factors are still uncorrelated, but they have different variances. For the factor loadings matrix, **A**, we have that

$$\mathbf{A}'\mathbf{A} = \mathbf{U}_k'\,\mathbf{U}_k = \mathbf{I};$$

that is, the columns of **A** are orthonormal. Furthermore,

$$\mathbf{S} = \mathbf{Y}'\mathbf{Y}\ \mathbf{AF}'\mathbf{FA}' = \mathbf{A}\Lambda_k\mathbf{A}'.$$

Assuming, again, that $N \geqq p$, one computes the k largest eigenvalues of **S**, Λ_k, and the corresponding eigenvectors, $\mathbf{U}_k$. Then, $\mathbf{A} = \mathbf{U}_k$ and $\mathbf{F} = \mathbf{YA}$.

This is the solution of interest in the present paper. No scaling of columns of $\mathbf{U}_k$ is needed. In the studies accounted for here, $k = p$.

In solution one, all the factors are uncorrelated and have the same variance, one, so that the columns of **A** are directly comparable. In solution two, the factors are also uncorrelated, but they have different variances, so the columns of **A** are not directly comparable.

Clearly, the approach offered by solution one does not lead to interest in the smallest roots and their eigenvectors, as these would tend to be looked upon as not contributing significant information. In my opinion, based on empirical studies, for the elements of the last eigenvector to be useful, the sample size must be large, with N around 100 objects for 10 variables. This is, naturally, not a hard and fast rule, as the important point is that the variances and covariances should be stable. Moreover, if there is much random variation in the variances and covariances, this, coupled with the rounding errors accumulating in most methods of extracting eigenvalues and eigenvectors will make for undue fluctuations in the elements of the last eigenvector.

Some Examples

The only way in which support for the ideas expressed in the foregoing section can be obtained is by empirical methods. I discuss now a few examples, but it should be understood that these are no more than a randomly chosen set of studies and no claim is made for their being the best possible material for illustrating the problem at hand.

Danian planktonic foraminifers. — Malmgren (1974) published an extensive account of the morphometry of some species of Danian foraminifers from Southern Scandinavia. The variables

Table 1. Sixth eigenvector for planktonic foraminifers.

	Subbotina pseudobulloides		*Globoconusa daubjergensis*
x_1	0,82	0,80	0,70
x_2	−0,39	−0,47	−0,50
x_3	0,10	0,08	0,14
x_4	−0,32	−0,33	−0,43
x_5	0,05	0,08	0,21
x_6	−0,26	−0,12	−0,11
var %	1,2	1,2	0,6
N	100	150	100

he considered are: length of test (x_1), width of test (x_2), height of test (x_3), width of final chamber (x_4), width of penultimate chamber (x_5) and width of pre-penultimate chamber (x_6).

In Table 1, the elements of the last eigenvector of the covariance matrix of logarithmically transformed variables are listed for the species *Subbotina pseudobulloides* and *Globoconusa daubjergensis*. The close agreement in elements for the two samples of *S. pseudobulloides* needs no comment. For purposes of comparison, a sample of another planktonic foraminifer, *G. daubjergensis*, was included. That both are coiled is reflected in the relative magnitudes of the elements of their last eigenvector, but there are certain clear differences. This could possibly be taken as an indication that the eigenvector of the smallest component may be taxonomically useful, in that it allows direct comparison between proportions of the variables in a constant, or almost constant, linear relationship.

A Maastrichtian benthic foraminifer. — The covariance matrices of fifty three samples of the Maastrichtian (Late Cretaceous) bolivinid foraminifer *Afrobolivina afra* Reyment were subjected to principal components analysis. The following nine variables were measured on each of 590 tests: x_1 = length of test, x_2 = maximum width of test in the plane of biseriality, x_3 = width of last chamber, x_4 = height of last chamber, x_5 = height of second last chamber, x_6 = diameter of proloculus, x_7 = breadth of test at right angles to plane of biseriality, x_8 = width of aperture, x_9 = distance of aperture from edge of second last chamber.

Table 2. Fourth eigenvectors for grasshoppers

	Kinnekulle		Öland		Gotland	
	males	females	males	females	males	females
x_1	0,92	0,83	0,78	0,80	0,80	0,82
x_2	−0,11	−0,52	−0,55	−0,49	−0,60	−0,47
x_3	−0,37	−0,20	−0,27	−0,35	0,02	−0,33
x_4	−0,07	−0,03	0,07	0,01	−0,04	0,02
var %	3,89	3,99	4,31	2,92	4,41	3,15
N	110	101	120	117	128	95

The eigenvectors for 21 samples with very small eigenvalues (the smallest principal components) are completely dominated by an equally weighted negative bipolarity between variables x_1 and x_2, suggesting that these two variables occur in a constant ratio to each other. A large, and probably biologically significant, part of the material shows little variability in the relationship between the maximum width of the test and the width of the last chamber.

Geographical variation in recent grasshoppers. — Three geographical isolates of *Omocestus haemorrhoidalis* occur in Sweden (Reyment, 1969). In Table 2, I have listed the eigenvectors of the smallest eigenvalues for males and females. The variables are: length of hind femur (x_1), pronotal length (x_2), elytron length (x_3), and the least width between ridges on the pronotus (x_4).

In this case, the least eigenvalue is quite large and it can hardly be expected that the corresponding eigenvector will indicate a non-varying relationship between the variables. Perusal of Table 2 indicates that the tendency is there, but it is obscured by noise.

Mineral chemistry. — Saxena (1969) studied silicate solid solutions and geo-thermometry from the point of view of the distribution of iron and magnesium between co-existing garnet and biotite, using principal component analysis. Ninety-three samples of rocks formed at various pressures and temperatures were analysed for the following variables:

x_1 = the distribution coefficient on a one cation exchange basis

$$k_{D_{Fe}}^{\text{Bi-Gar}} = \frac{x_3(1-x_2)}{(1-x_3)x_2}$$

x_2 = Fe in garnet
x_3 = Fe in biotite
x_4 = Mn in garnet
x_5 = Ca in garnet
x_6 = Al^{iv} in biotite
x_7 = Al^{vi} in biotite
x_8 = Ti in biotite.

The principal components of the correlation matrix were extracted, with which, Saxena was able to distinguish between rocks of low and high metamorphic grade. The smallest eigenvalue ac-

counts for 0,32 % of the total variance. It is associated with the eigenvector

(−0,58; −0,66; 0,50; 0,09; 0,00; −0,01; −0,04; −0,15).

This variable can be interpreted as being an invariant relationship between the first three variables, to wit, $-0{,}53x_1 - 0{,}66x_2 + 0{,}50x_3$. What we have here is clearly the expression of the distribution coefficient k_D, a ratio expressing x_1 in terms of x_2 and x_3. In other words, the smallest principal component is a result of the derived variable, x_1.

REFERENCES

Hotelling, H., 1933: Analysis of a complex of statistical variables into principal components. *J. Ed. Psych.*, 24, 417—441, 498—520.

Gnanadesikan, R. and Wilk, M. B., 1969: Data analytic methods in multivariate statistical analysis. *Multivariate Analysis,* 2, Academic Press, NY, 593—638.

Gower, J. C., 1967: Multivariate analysis and multidimensional geometry. *Statistician*, 17, 13—28.

Jöreskog, K. G., Klovan, J. E., Reyment, R. A., 1976: *Geological Factor Analysis.* Elsevier, Amsterdam.

Malmgren, B. A., 1974: Morphometric studies of planktonic foraminifers from the type Danian of Southern Scandinavia. *Stockh. Contr. Geol.*, 29, 1—126.

Reyment, R. A., 1969. Some case studies of the statistical analysis of sexual dimorphism. *Bull. geol. Instn. Univ. Upsala,* NS 1, 97—119.

Reyment, R. A., 1978: Интерпретация наименьшей главной компоненты. Исследования по математической геологии. Publ. Akad. Nauk, USSR, 163—167.

Saxena, S. K., 1969: Silicate solid solutions and geothermometry. *Contr. Min. Petrol.*, 22, 259—267.

Printed by Ekenäs Tryckeri Ab, Ekenäs, Finland April, 1979.