Bull. geol. Instn Univ. Upsala N.S. 2: 8, 67–72, 1970
(Symposium on Biometry in Paleontology)

EIGEN-THEORY IN NUMERICAL TAXONOMY

R. A. Reyment

Paleontological Institute, University of Uppsala

Abstract. The numerical-taxonomic method of principal coordinate analysis is discussed in relation to the general problem of quantitative classification. The method is illustrated by a biologic example involving the ostracod genus *Buntonia* and a sedimentologic example treating the classification of interstitial environments.

INTRODUCTION

Recently there has been a considerable growth of interest in the possibility of numerical methods as a major aid to classification of, usually, biological individuals. A detailed review of several methods of numerical analysis is given in Sokal & Sneath (1963) as well as a review of the historical development of the subject. All of these methods begin with a multivariate sample but each individual may be, though not necessarily, from a different biologic population. The variates may be quantitative, qualitative, dichotomous or a mixture of these, hence, product-moment correlations may not be appropriate. Generally a so-called association matrix provides the working basis and in the present study that of Gower (MS) has been used.

There is a section of the subject of Multivariate Statistical Analysis which has always been titled "Classification". This occurs in conjunction with what is commonly referred to as (Multiple) Discriminant Function Analysis. As has been pointed out by others, this is really not a matter of classification *per se*, but rather one of *identification*. At least, this is the interpretation that must be given as soon as the technique is applied to taxonomic situations.

A moment's reflection will disclose why this must be so. If we consider the statistical discrimination problem for the case of two populations, compatible with respect to k variables (populations π_1 and π_2) and we have an observational vector,

$$X = (x_1, \ldots, x_k),$$

it is required to find to which of populations π_1 and π_2 the vector, X, belongs. This is thus a question of IDENTIFICATION with either of π_1 and π_2. If this is not the basic form of model desired, then it is wrong to employ the DISCRIMINANT FUNCTION in this connexion. It is well known, that visual methods of traditional stamp in classification are deeply influenced by subjectivity. It is enough to compare contemporary publications in one's own field of research. This may be called the *first category of classification problem*. That is, where one already has a form of taxonomic classification in existence, but it suffers from defects of several kinds, introduced in varying degrees by subjectivity. It is desirable to be able to better this, by some means or other, in order to introduce the element of REPEATABILITY. Thus, if our model is a reliable one, it should be possible for any other person in the field of research to be able to take the material and, without *a priori* knowledge of what has been drawn in the way of classificatory conclusions, end up with identically the same result.

The second kind of classification problem occurs mainly in non-biologic sciences. Problems embracing topics stretching from Permian cyclothems, the classification of mineral resources, to the classification of archeologic objects. Here, the application is concerned with developing a classification scheme, on quantitative grounds, on little or no *a priori* information. From a cursory perusal of the literature falling into the second kind of problem, it seems as though the models of the biologists have been employed with little or no modification; clearly the resolution can only be as good as the underlying model.

NON-STATISTICAL NUMERICAL TAXONOMY

Modern numerical taxonomists, such as J. Rohlf and R. Sokal, have departed largely from the realm of mathematical statistics in developing their subject

owing, largely, to the great number of special problems and difficulties that arise when a statistical basis is adopted. This is, in my opinion, not always such a good idea, as this approach is frequently adopted to get around something unpleasant in the data.

As is well known, the concept of numerical taxonomy appears to have been born with M. Adanson, the French naturalist, who formulated the concept in conjunction with his taxonomic studies on the Recent marine molluscs of Senegal, West Africa.

One of his postulations was that of the *principle of equal weighting* of all the characters selected by the zoologist as diagnostic of his material (the question of what is to be regarded as diagnostic is, of course, a moot point). This principle of equal weighting would appear to be the one most widely practised among numerical taxonomists of today. The obvious logic employed in support hereof is, that subjective elements would be introduced in that the quantitative zoologist would be exercising personal opinions and prejudice in the choice-making procedures if equal weighting were not resorted to. However, if one examines the logic behind the concept of the SIMILARITY COEFFICIENT of the numerical taxonomists, in the garb presented by Sokal & Sneath (1963), it soon becomes apparent, that this is an area in which personal opinion is permitted considerable rein and in actual fact, it transpires, that some of these similarity coefficients may only owe their unlikeness to some form of character-weighting or other.

What does the critic of equal character weighting have against it? The most ready-to-hand complaint would seem to be, that any definition of what is a diagnostic and useful character (often termed a "unit character") lies in the mind of the person carrying out a particular study and, of necessity, will be subjectively flavored. Each worker will view a certain situation in a different light from another. Hence, the claim of objectivity in conjunction with the principle of equal weighting is one that should invite a certain measure of sceptisim.

It is natural enough to ask, whether it might not be possible to produce a character-weighting coefficient, which will in some manner compensate for the lack of pertinence in a chosen character. The *generalized statistical distance* of P. C. Mahalanobis provides a method of character-weighting, whereby the introduction of a new character to a set of characters causes little or no change in the "distance", if this character (characters) is (are) strongly correlated with characters existing in the set. We may state this in other terms, notably, that if a character conveys no new information for separating between two samples (or populations), the pertinent elements of the inverse covariance matrix of the quadratic form of the generalized distance will be very small, to use non-precise language.

THE MEASURE OF ASSOCIATION

The general applicability of the D^2-method is, to a degree, limited by the difficulty of satisfactorily using it in conjunction with discrete characters, and there are some other problems, such as the *a priori* establishing of the basic groups. Suggestions have been made, that a possible approach is by means of an information-theoretic quantity. This requires the estimation of prior probabilities for character states, these determining the weights of the states. Other opinions with respect to the *weighting dilemma* are represented in the literature. I mention this in order to bring out the fact, that several points of view are developing in the non-Adansonian sphere.

Observations on the concept of cluster analysis in numerical taxonomy have been most recently given by Gower (1967). It is not here proposed to enter into a review of similarity coefficients but it does seem desirable to mention that Gower (MS) has proposed a useful measure, which leads to a positive definite similarity matrix. For quantitative characters with values $x_1, x_2, \ldots x_n$ of character k for the total sample of n individuals Gower's coefficient is defined as

$$s_{ijk} = 1 - |x_i - x_j| / R_k. \qquad (1)$$

Here R_k is the range of character k. The matrix S_{ij} with elements s_{ijk} ranges in value between 0 and 1. This coefficient of similarity has been used in the present study and has therefore been mentioned in detail, although there are other coefficients equally worthy of consideration (cf. Sheals, 1964).

PRINCIPAL COORDINATE ANALYSIS

Eigenanalysis of a matrix of associations

A useful way of considering classification procedures is that embodied in the concept of Numerical Taxonomy in the form presented by, for example, Sokal & Sneath (1963). An objection to this approach has been that the methods and procedures of numerical

taxonomy lack, to a considerable extent, a firm mathematical basis. Gower (1966) has made an encouraging move in a direction towards strengthening the foundations of the subject and it is to be expected, that more mathematical statisticians will be drawn towards this area of research. As an accessory to the method of principal components (an *R* technique), we shall consider the procedure termed "principal coordinates" by Gower (1966, p. 137).

In a numerical taxonomic study, there will be measurements on p variates for each of n individuals. The interrelationships between these variates will be estimated by means of some form of coefficient of association, a_{ij}, between all pairs of individuals. These form the "association matrix" A.

We consider the symmetric matrix A of order n. The eigenvalues of A are $\lambda_1, \ldots, \lambda_n$ and its eigenvectors are $b_1, b_2, \ldots b_n$. These form the matrix B of order n. In applying this in numerical taxonomy, one takes the elements of the ith row as the coordinates of a point, T_i, in n-space. The distance, d_{ij}, between P_i and P_j is then given by

$$d_{ij}^2 = \sum_{r=1}^{n} b_{ir}^2 + \sum_{r=1}^{n} b_{jr}^2 - 2\sum_{r=1}^{n} b_{ir}b_{jr}. \tag{2}$$

Thus, for rows 2 and 3

$$d_{23}^2 = \sum_{r=1}^{n} b_{2r}^2 + \sum_{r=1}^{n} b_{3r}^2 - 2\sum_{r=1}^{n} b_{2r}b_{3r}. \tag{3}$$

If the eigenvectors of A are normalized so that the sums of squares of their elements are equal to the corresponding eigenvalues,

$$b_{ir}^2 = \lambda_r,$$

then

$$A = b_1 b_1' + b_2 b_2' + \ldots + b_n b_n' \tag{4}$$

and

$$d_{ij}^2 = a_{ii} + a_{jj} - 2a_{ij}. \tag{5}$$

By this means, one may represent a multivariate sample of size n as points $T_1, \ldots T_n$ in a Euclidean space. The relationship between the eigenvalues and eigenvectors of the association matrix A are indicated in Table I.

Gower (1966) has demonstrated that it is legitimate to use the methods of principal components as a Q technique on the coordinates of the T_i to find the best fit in fewer dimensions. In accord with what is usually observed in principal components, it may be expected that a good representation of the set of points may be obtained in a reduced number of dimensions when some of the eigenvalues are small. That is if, say λ_r is small, the contribution of $(b_{ir} - b_{jr})^2$ to the distance between T_i and T_j will be small. If λ_r is large but the b_{ir} corresponding to it are not greatly different then $(b_{ir} - b_{jr})^2$ will be small. Hence, the only coordinates supplying much to the distances are those which display a wide range of variation in the elements of the eigenvectors and which are associated with a large eigenvalue. In common with what is found in principal component analysis, the distances may often be adequately expressed by two or three vectors.

Table I. *Eigenvalues and eigenvectors of association matrix.*

		Eigenvalues			
		λ_1	λ_2	...	λ_n
		Eigenvectors			
Point	T_1	b_{11}	b_{12}	...	b_{1n}
	T_2	b_{21}	b_{22}	...	b_{2n}
	.	.	.		.
	.	.	.		.
	.	.	.		.
	T_n	b_{n1}	b_{n2}		b_{nn}
Centroid	$\bar{T}$	$\bar{b}_1$	$\bar{b}_2$		$\bar{b}_n$

A further analog with *R*-technique principal components is that the elements of the first eigenvector may often be found to have similar elements, relating to the mean value of all the elements of A. This mean value is not important to the problem at hand, as the distances are invariant for any constant added to A.

The addition of such a constant will, however, result in different coordinate values T_i and different eigenvalues and these new points are an orthogonal transformation of the original set after a change of origin. Interest naturally attaches to determining which transformation gives the best fit with a reduced number of coordinates.

This mean value is unimportant in this connexion as the addition of any constant to the association matrix A leaves the distance between T_i and T_j invariant. Gower (1966, p. 330) adjusts for the means in the following way. It is always possible to adjust matrix A so that it has a zero eigenvalue without altering the distance between T_i and T_j, for if $\bar{a}_i$ is the mean value of the ith row or column of A, and $\bar{a}$ is the overall mean, a matrix α_{ij} may be defined in terms of the elements

$$\alpha_{ij} = a_{ij} - \bar{a}_i - \bar{a}_j + \bar{a}. \qquad (6)$$

Inasmuch as every row and column of matrix α_{ij} sums to zero, α_{ij} has a zero eigenvalue.

Computer program. The computer program used for the calculations first forms the association matrix A, which is then transformed to matrix α_{ij} by equation (6). The eigenvalues and eigenvector of matrix α_{ij} are computed and each eigenvector is scaled so that its sum of squares is equal to the corresponding eigenvalue.

The ith row of Table I represents the coordinates of a set of points T_i whose distances apart are given by the best approximations to $(a_{ii} + a_{jj} - 2a_{ij})^{\frac{1}{2}}$ in the chosen number of dimensions.

The computed coordinates are plotted by a simple plotting subroutine included in the program. Storage space is always a troublesome matter in a program of the kind considered here. The writer's program will take 95 individuals for any number of variables. The *Buntonia* example was run for 70 individuals and took about 6 minutes on a CD 3600.

In summary, the method of principal coordinate analysis finds the coordinates of each individual of a sample, referred to principal axes, which preserve the distances, suitably defined, between the individuals.

EXAMPLES

Example 1. Subdivision of a biologically homogeneous sample of *Buntonia olokundudui* Reyment and Van Valen. The material (70 individuals) of *Buntonia* is from the Niger Delta, West Africa. The four variables are: length of carapace, and the numbers, respectively, of the anterior, posterior and lateral spines. The Gower association matrix is of order 70. The eigenvalues and eigenvectors of the matrix given by (6) were extracted by the Jacobi procedure, which for the program used required 12979 iterations. The sum of the eigenvalues is 16.585 and the sum of the first two eigenvalues, approximately 11; these account for most of the "variation" and a reasonably efficient set of coordinates should be provided by the corresponding eigenvalues (4). The most important eigenvalues are shown in Table II.

Table II. *The largest eigenvalues of the adjusted association matrix.*

Number	Eigenvalue	Number	Eigenvalue
1	9.4578	7	0.5223
2	2.3640	8	0.4349
3	1.7714	9	0.3393
4	1.0584	10	0.3064
5	0.7995	11	0.1966
6	0.5629	12	0.1624

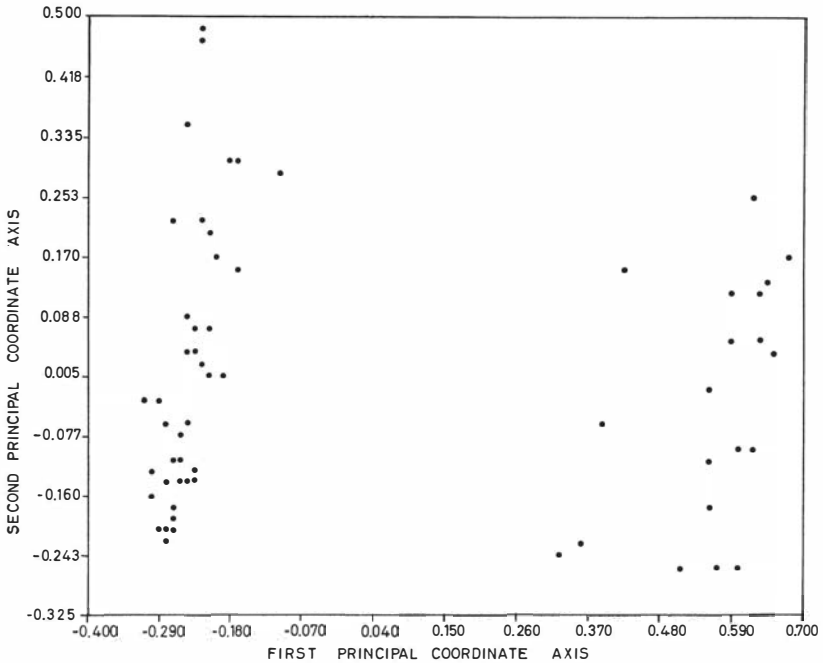

Fig. 1. Plot of first two principal coordinates for *Buntonia.*

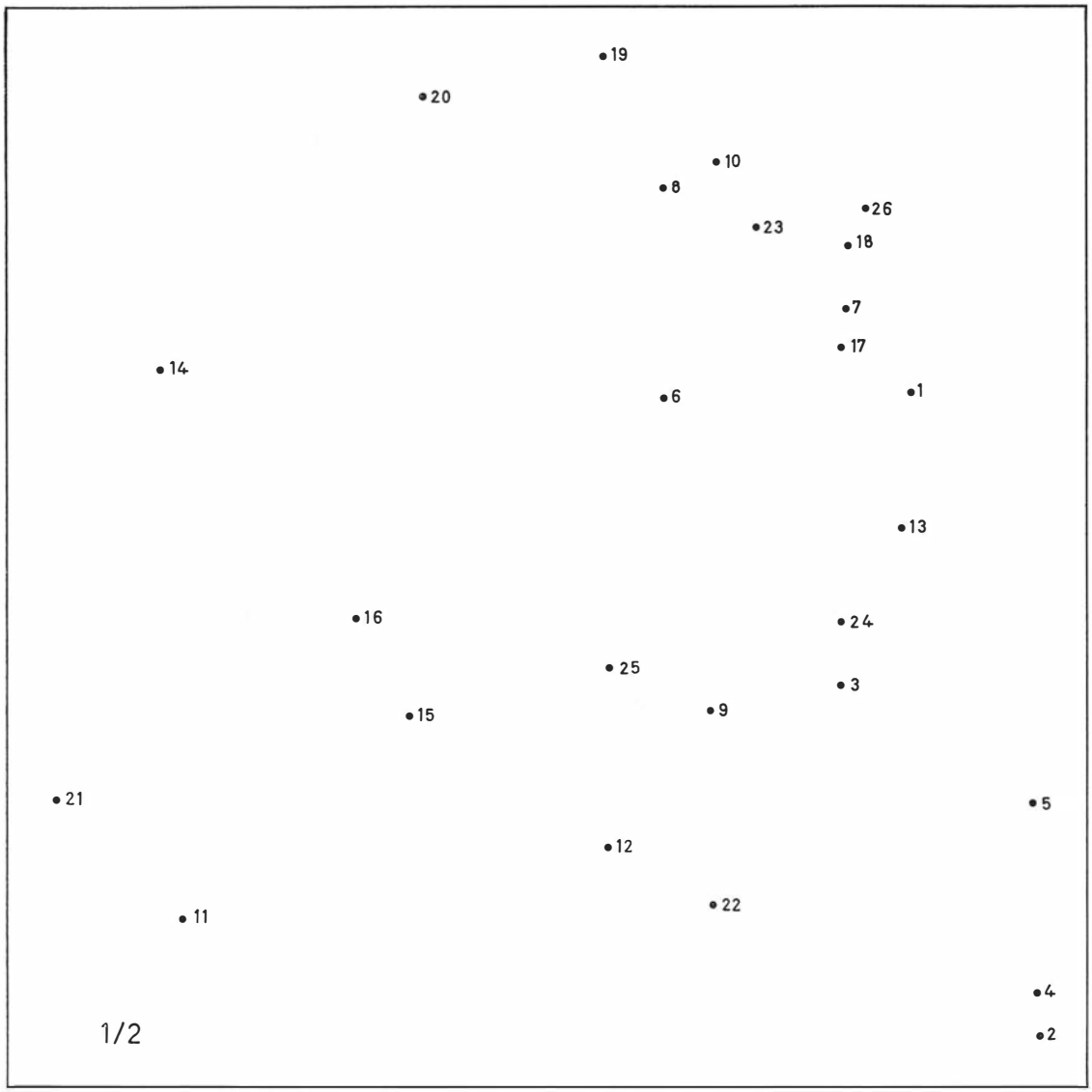

Fig. 2. Plot of first two principal coordinates for sedimentary data.

Already the fourteenth eigenvalue is about 0.1 and the successful concentration of the association information to the first few eigenvalues and the rapid taper off of the subsequent roots to near-zero values suggests that the classificational efficiency in this problem may be high and that there is a rather high degree of association between individuals. The plot of the first two coordinates (cf. Table I) is shown in Fig. 1; the coordinates are listed in Table III. There is a clear subdivision into three groups, which is a reflection of the discrete nature of the variation in the species of *Buntonia*. One might intuitively expect such a breakdown of data of this kind; however, the useful feature yielded by the principal coordinate analysis is that the morphologic categories with the most in common are placed together. The elongated forms of the distributions are largely a result of the several growth stages forming the material. The analysis of this problem may be considered to have given a satisfactory result.

Example 2: Sedimentary interstitial environmental categories. A survey of bottom sediments in the Niger Delta, made by the writer in April–May, 1966, suggested the possibility of several environmental categories with respect of the interstitial environment

Table III. *First two sets of principal coordinates for* Buntonia.

1	2	1	2	1	2	1	2	1	2
0.64	0.05	−0.24	−0.14	−0.27	−0.06	−0.26	−0.17	−0.29	−0.15
−0.20	0.20	−0.24	−0.14	−0.23	−0.14	0.43	0.16	−0.28	−0.20
0.57	−0.02	−0.26	0.21	−0.18	0.01	0.63	−0.10	−0.21	0.21
−0.26	−0.12	−0.25	−0.08	0.38	0.26	0.60	0.12	0.66	0.04
0.36	−0.22	−0.17	0.31	0.40	−0.06	−0.24	0.09	−0.24	−0.14
0.60	−0.09	−0.27	−0.22	−0.27	−0.19	0.60	0.05	−0.23	−0.13
0.61	−0.26	−0.21	0.47	−0.22	0.02	−0.25	−0.14	0.56	−0.11
−0.25	−0.14	−0.24	−0.06	−0.21	0.08	−0.30	−0.12	−0.10	0.28
−0.20	0.01	−0.23	0.04	−0.31	−0.03	−0.28	−0.15	−0.17	0.15
−0.27	−0.22	−0.28	−0.03	0.62	0.26	0.68	0.17	−0.20	0.16
−0.23	−0.14	−0.24	0.08	0.57	−0.26	−0.28	−0.21	−0.22	0.25
−0.25	−0.12	−0.22	0.22	0.65	0.13	−0.17	0.30	−0.23	0.08
−0.22	0.02	−0.21	0.48	0.64	0.13	−0.24	0.35	0.34	−0.23
−0.26	−0.20	−0.28	−0.21	−0.24	0.04	0.56	−0.17	0.52	−0.26

and mineral constituents of the sediment. The principal coordinate analysis did not, however, disclose the existence of significant groupings in the data, the suggestion being that there are no sharp boundaries. The eigenvalues decrease only slightly from root to root suggesting that there is much near random variation in the material. The plot of the first two coordinates is shown in Fig. 2; the lack of a tendency for points to cluster is clearly discernible.

Sommaire. La méthode de taxonomie numérique dans l'analyse des coordinés principaux est discutée en sa relation au problème général de classification quantitative. La méthode est illustrée par un exemple pris de la biologie où il est question du genre ostracode *Buntonia* et un exemple pris de la sédimentologie, traitant de la classification des milieux interstitiels.

REFERENCES

Adanson, M. 1757. *Histoire naturelle du Sénégal. Coquillages. Avec la relation abregée d'un voyage fait en ce pays, pendant les années 1749, 50, 51, 52 et 53*. Bauche, Paris.

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

— 1967. A comparison of some methods of cluster analysis. *Biometrics* 23, 623–637.

Sheals, J. G. 1964. The application of computer techniques to Acarine taxonomy; a preliminary examination with species of the *Hypoaspis-Androlaelaps* complex (Acarina). *Proc. Linn. Soc. Lond.* 176, 11–21.

Sokal, R. S. & Sneath, P. H. A. 1963. *Principles of Numerical Taxonomy*. Freeman, San Francisco–London.