MULTIVARIATE ANALYSIS IN ECOLOGY

R. M. CASSIE

Zoology Department, University of Auckland.

Ever since the discipline was first established, ecologists have emphasised the extreme complexity of ecology. This complexity has frequently rendered the interpretation of field data almost impossible, or, when interpretation has succeeded, it has usually encompassed the more obvious phenomena, and one suspects that, as with an inefficient gold dredge, much of value is left behind in the "tailings". Much (though certainly not all) of the necessary mathematical theory needed to extract the lost information has long been available, but has either been unknown to the ecologist, or so time-consuming in application that its use was impracticable. With the advent of the high-speed computer, the logistic difficulties, at least, have been removed; and the power and elegance of multivariate techniques have been demonstrated in ecology by pioneering works such as those of Goodall (1954) and Williams and Lambert (1959), although Goodall's work was in fact carried out on a Facit hand calculator. At the same time, the application of mathematical theory to ecological situations is not entirely clear-cut, and many of the methods and concepts are still subject to controversy.

Over the past few years I have been developing a series of computer programmes designed to interpret tables of ecological data, and what follows is an attempt to develop from relatively simple statistical concepts an explanation, if not a justification, of the techniques I am using. In doing so, I exhibit some of my own preferences; and I must admit that some of these would undoubtedly be challenged by other workers. I have preferred the relatively simple methods such as principal component analysis to the more sophisticated "factor analysis" methods which have their origin largely in psychological research. For example, I can see no particular reason why a given number of variates should have their basis in a smaller number of "factors" — surely the distribution of, say, p different species will be determined by *at least* p independent factors; and the most one can hope for is that as much as possible of the distribution pattern can be accounted for by as few as possible main factors. I have a similar, possibly intuitive, suspicion of "simple structure" concepts (Thurstone, 1947), because I cannot believe that ecology (or, for that matter, psychology) has a simple structure. Rather similar objections have been raised by Seal (1964, pp. 167 *et seq.*).

I also prefer to deal with continuously variable data (measurements or moderately large counts) rather than binary (presence or absence) data; though I accept that the latter are much easier to collect and may sometimes be the only form possible. Almost all the statistical theory for dealing with continuous data is based on the multivariate normal distribution. Methods do exist which claim to be independent of distribution but these claims should be treated with suspicion — they are often the same methods with embarrassing sections of the theory conveniently omitted. It is not true, as some authors have implied (often ambiguously), that methods such as principal component analysis are independent of distribution. Provided one is interested only in the properties of the sample, almost any manipulation of the data may be legitimate. However, the ecologist is interested not in the sample but in the population it represents, and no valid estimate of the *population* principal components can be made unless the distribution of the variates is specified.

Closely linked with normality is linearity. It is much easier to construct and analyse linear relationships of the type:

$$Y_i = k + \beta X_i \qquad (1)$$

where i refers to the ith observation, X and Y are variables, and k and β are constants. Often we find biological relationships of the type:

$$Q_i = K \,.\, \exp(\beta X_i)$$

where, for example, Q is the rate of reproduction, metabolism, etc., and X is the temperature. This is easily reduced to a linear form by transforming. Putting: $Y_i = lnQ_i$, $k = lnK$, we return to the form of (1). If X is normally distributed, so is Y, but untransformed Q is strongly skewed and will lead to considerable statistical difficulties, not only (as some have implied) in tests of significance but also in estimation of parameters. One of the simplest statistical relationships which can exist between two variables is linear regression:

$$Y_i = k + \beta X_i + \varepsilon_i \qquad (3)$$

where an element of uncertainty has been introduced into (1) by adding the error term, ε_i. Given a set of corresponding values of X_i and Y_i, elementary statistical methods allow us to estimate k, β, and (for the ith $X:Y$ pair) ε_i. Notice that in (3) the properties of normality and linearity are entirely separate. Y_i *is* linearly related to X_i, but we need know nothing about the distribution of either. The only true variate (=random variable) is ε, which is assumed to be normally distributed. Thus regression is not bivariate and multiple regression is not multivariate as may be seen from the equation:

$$Y_i = k + \Sigma_j \beta_j X_{ij} + \varepsilon_i \qquad (4)$$

Before investigating true multivariate situations, it is convenient to modify (3) to:

$$y_i - \varepsilon_i = \beta x_i \qquad (5)$$

For Y_i and X_i we have substituted y_i and x_i, which have zero means (taken over all values of i), i.e.:

$$y_i = Y_i - \bar{y}$$
$$x_i = X_i - \bar{x}$$

This enables us to eliminate the constant, k. Also, ε_i is shifted to the left-hand side of the equation to emphasise that the error is attached to Y rather than to X.

Although most statistical texts emphasise that the independent variable, X, in regression must be measured without error, it is not always made clear what this statement means. Error in this sense is not only a matter of precise measurement. If, for example, X is a temperature and Y the abundance of an organism, the thermometer may be accurately and correctly read, but the temperature may still not be completely appropriate to the situation. The temperature determining abundance of organisms is probably the *average* temperature over the last day, month or year, of which the temperature read at the time of sampling is only an estimate. Equation (5) then becomes:

$$y_i - \varepsilon_i = \alpha(x_i - e_i) \qquad (6)$$

where ε and e are normal variates. This is a truly bivariate situation, and one which cannot be solved without further information beyond that given by the table of Y_i and X_i. The regression coefficient, β, can be (and often is) an entirely erroneous estimate of the "underlying" coefficient, α. To estimate α, ε, and e, we would need to know the relative variances of ε and e, i.e. the relative amount of error contributed by the Y and X observations. One possible solution might be to set up a controlled laboratory experiment in which organisms were raised at constant temperatures, thus making e_i and its variance zero. However, apart from being expensive and time consuming, it is even doubtful whether this experiment would reproduce any true natural situation, because no organism in nature ever lives in such a constant environment.

An idealised (though doubtless oversimplified) equation for the relationship between the abundance of a number of kinds of organisms and properties of their environment might be:

$$\Sigma_j \alpha_j (y_{ij} - \varepsilon_{ij}) = \Sigma_k a_k (x_{ik} - e_{ik}) \qquad (7)$$

Again it is implied that the ys and xs are linearly related one to another, and that the εs and es are normally distributed and uncorrelated with the xs and ys (though not necessarily independent of one another, since it may be possible to resolve the error terms of one equation into further equations of similar nature). In practice, the segregation of linear and normal components is a formidable, if not impossible task, so that we resort to a simpler, but more restrictive model where the ys and xs are not only linearly related, but are also normal variates, which can be partitioned into two components, also normal. Thus, for example:

$$y_{ij} = v_{ij} + \varepsilon_{ij} \qquad (8)$$

The assumption of normally distributed ys and xs tends to be restrictive in some ecological situations. Even after applying appropriate transformations, it is commonly found that a "variate" has, not a single normal distribution, but two or more such distributions intermingled. If one imagines a number, say p, of variates plotted one against another in "p-dimensional" space, the multivariate normal model calls for the points to form a cluster something like the plums in a pudding, each plum representing a sample (or quadrat, etc.) and with the greatest density near the centre of the pudding. In practice, one often finds not one but a number of such clusters. Sometimes the various clusters tend to overlap one another, and even the number present appears to vary according to the angle from which the pudding is viewed. It is, of course, impossible to view anything in 4 or more dimensions, but it is relatively easy to rotate the pudding (by mathematical means) to bring any two or three dimensions into real space for observation. A computer with a high-speed line printer can quite easily be programmed to produce scatter diagrams in two dimensions, and if the most informative dimensions are chosen the structure of the pudding can be studied. Programmes have also been produced for

plotting 3-dimensional scatter diagrams (viewed with a stereoscope — Rohlf, 1968), but these require the use of a plotter which is relatively expensive both to purchase and to operate.

Since we cannot derive a rigorous solution even for the bivariate case, it is not surprising that multivariate analysis produces an enormous number of difficulties and differences of opinion among its exponents. Probably there is no one ideal solution for any given situation, and it is even difficult to make an objective choice between one solution and another. My own opinion is that a solution which is acceptable to a competent ecologist (if only on intuitive grounds) is better than one which is not. This, of course, lays me open to the criticism that "He only accepts the analysis if it agrees with what he expected in the first place". This is difficult to answer, but the ecologist often has to tread along the precarious path between subjectivity and objectivity. At the same time, I would not advocate abandoning statistical rigour altogether. When appropriate, variates should be transformed to approximate as closely as possible to linearity and normality, and of the two criteria it seems likely that (at least in ecology) linearity is the more important. Tests of significance should be applied where they are available, but at the same time remembering that dimensions discarded as "non-significant" are not necessarily non-existent — they are merely dimensions about which we do not have sufficient information.

So far, of the various techniques I have used, the one which seems to give the most "meaningful" information is principal component analysis derived from the correlation matrix of the logarithms of the species counts. This has the effect of producing a comparable function to the left-hand side of (7). There are as many vectors as there are species, but the number may be substantially reduced by deleting those with the smaller variances (or eigenvalues), so that the amount of information so discarded is minimal. Tests of significance are not necessarily helpful in deciding how many vectors to discard, because in a large set of samples it is quite usual for all eigenvalues to be formally significant. Such an analysis does not take into account the known properties of the environment, i.e. the xs in (7), but one may perhaps assume that the organisms themselves "know" more about the kind of environment they favour than does the ecologist. They display this "knowledge" by arranging themselves into "communities" which reflect common preferences for certain types of environment (recognising, however, that not all ecologists will accept such a definition of a community). Each vector now represents a community, and the members of the community are those species which have their largest element in the vector of that community.

TABLE 1. *Eigenvalues and eigenvectors for Karore Bank.*

eigenvalue	3.62	1.80	1.36
Chione stutchburyi	**0.83**	0.06	−0.10
Macoma liliana	**0.80**	−0.38	0.07
Amphidesma ventricosum	**0.59**	0.13	0.35
Nucula hartvigiana	0.55	−**0.65**	0.17
Memiplax hirtipes	0.08	−**0.67**	−0.50
Cyclomactra discors	−0.20	0.11	−**0.54**
Pectinaria australis	−0.32	−0.12	−**0.58**
Solemya porkinsoni	−**0.42**	−0.30	0.33
Soletellinia sillgua	−**0.44**	−0.40	0.20
Leptomya retiaria	−**0.49**	−0.46	0.04
Halicarcinus cookii	−**0.62**	−0.36	0.32
Owenia fusiformis	−**0.71**	0.37	0.22

Tables 1 and 2 represent a principal component analysis of the abundance of invertebrates on two intertidal mudflats. On the basis that each species contributes one unit of variance, the eigenvalues (sometimes also called latent roots) at the top of each table are the variances which have been accounted for by each of the three vectors retained. For example, on Karore Bank, the first vector accounts for $3.62/12 = 30\%$ and the first three vectors for $(3.62 + 1.80 + 1.36)/12 = 56.5\%$ of the information contained in the species-by-species correlation matrix. Thus these three mathematically defined communities (if we are permitted the term) describe the sample rather better than half as well as the original 12 species. Further, since only three roots were statistically distinguishable (or significant), even if we had retained more vectors, these would have given us no further information about the population (as opposed to the sample).

TABLE 2. *Eigenvalues and eigenvectors for Hobson Bay.*

eigenvalues	4.36	1.81	0.81
Chione stutchburyi	**0.91**	0.24	0.02
Nucula hartvigiana	**0.91**	0.20	−0.03
Cominella adspersa	**0.89**	−0.19	0.03
Zeacumantus lutulentus	**0.80**	−0.27	−0.19
Zediloma subrostrata	**0.75**	0.20	0.51
Macomona liliana	**0.71**	−0.64	0.04
Helice crassa	0.04	**0.90**	0.28
Amphibola crenata	−0.43	−0.57	**0.66**

The vector (column of figures) below each latent root is the set of coefficients, α, (7) multiplied by the appropriate standard deviation (square root of the eigenvalue). This does not affect the properties of the vectors, but makes vector elements directly comparable along any given row. The largest element in each *row* is shown in bold face, so that the bold elements in each column define the members of the community. The first vector for Karore Bank has the interesting feature of having large elements both with positive and with negative signs. We find that the positive and negative signs refer respectively to species favouring coarse and fine sediments. We must therefore consider them as forming two negatively correlated communities. Although both react to the same environmental factors, they react in different directions. Notice that there is nothing to prevent a species from belonging to two communities, and in fact *Nucula* in Table 1 comes close to doing this — it has just escaped being classified with the three other lamellibranchs, *Chione*, *Macomona* and *Amphidesma*.

Allowing for some element of subjectivity, this community diagnosis agreed well with what would have been predicted by two ecologists (Prof. J. E. Morton and Dr. M. C. Miller) familiar with these soft short associations (see Cassie and Michael 1968, for further details). However, a more objective test will be to conduct further surveys and analyses for other mud-flats with similar fauna and check for consistency of diagnosis. So far, only two such surveys have been made, both of them fairly small. At Hobson Bay (Table 2) we find a similar coarse sediment community including *Chione*, *Macomona* and *Nucula*. The fine sediment dwellers are less well represented, but the mud-snail *Amphibola* falls into this category. Another survey at Whangateau produced simpler results, scarcely requiring multivariate analysis, since the entire area consisted of relatively coarse sediments and, as might be expected, the three abundant species were members of the coarse sediment community, *Chione*, *Macomona* and (though restricted to the lower levels) *Amphidesma*.

Having defined the communities, the next step is to investigate what environmental conditions they are associated with. As yet, I have found no suitable procedure for estimating the *a*s in (7), but simple correlation coefficients seem to have a useful meaning. For example, for Hobson Bay, the first principal component has the following correlations with particle sizes taken in descending order of size:

0.79, 0.44, 0.38, 0.39 0.05, −0.50

Clearly, the six species comprising this community favour the coarsest sediment.

Another useful approach, which deals with *y*s and *x*s simultaneously, is canonical correlation. This is, in effect, a more general case of multiple regression and could be expressed by the equation:

$$\Sigma_j \beta_j y_{ij} = \Sigma_k b_k x_{ik} \pm \varepsilon_i \qquad (9)$$

Notice there is only one error term and that this is not specifically attached to either side of the equation. However, as in principal component analysis, it is assumed that the *y*s *and* *x*s are normally distributed, so that the model is not univariate but multivariate. As we might expect, since this is a technique closely related to regression, the βs and *b*s seem to have no useful meaning — in fact if we were to add or delete variates from either side of the equation, the values of corresponding coefficients would change quite capriciously. However, it *is* possible to estimate the coefficients of both sides of the equation, and in so doing to create a linear compound of the species (on the left) and another of the environment (on the right) the two compounds having maximum correlation. Thus, for example, for Hobson Bay, the highest canonical correlation (there are several possible) was 0.95.

A more complex situation occurred in a line of plankton samples taken along the long axis of Pelorus Sound (Cassie 1967). Here the measured environmental factors were temperature and salinity, but few of the species had any clear linear correlation with either. However, a very large correlation (0.97) was found in a function of the type:

$$\Sigma_j \beta_j y_{ij} = \Sigma_k b_k x_{ik} - \Sigma_k b'_k x_{ik}^2 \pm \varepsilon_i \qquad (10)$$

where x_{i1} is salinity and x_{i2} temperature. Thus, the relationship of organisms to environment was in this case not linear but quadratic, e.g. for a single species and a single property:

$$y_i = bx_i - b'x_i^2 \qquad (11)$$

Since this is the equation of a parabola with its apex directed upward, we may deduce that each species has an "optimum" for temperature and another for salinity, and that it declines in abundance on either side of this optimum. In this instance, "optimum" is probably a misnomer, since it is unlikely that any of the species is stenothermal of stenohaline within the range of the samples (0.33°C. and 0.5S). It is more probable that the physical properties are not so much an indication

of the preference of the organisms as an index of the oceanic water mass with which the species were associated when they entered the Sound. In order to illustrate the high degree of correlation between organisms and environment, the two sides of (9) have been evaluated and plotted (left-hand side against right) in Fig. 1. The numerical form of the equation is (for any given sample):

$$0.14y_1+0.03y_2-0.14y_3-0.16y_4-0.17y_5+0.02y_6-0.53y_7-0.19y_8+0.08y_9-0.14y_{10}-0.05y_{11}=1.42t-1.49s-1.25t^2+2.31s^2$$

where $y_1 \ldots y_{11}$ represent the species and t, s temperature and salinity respectively.

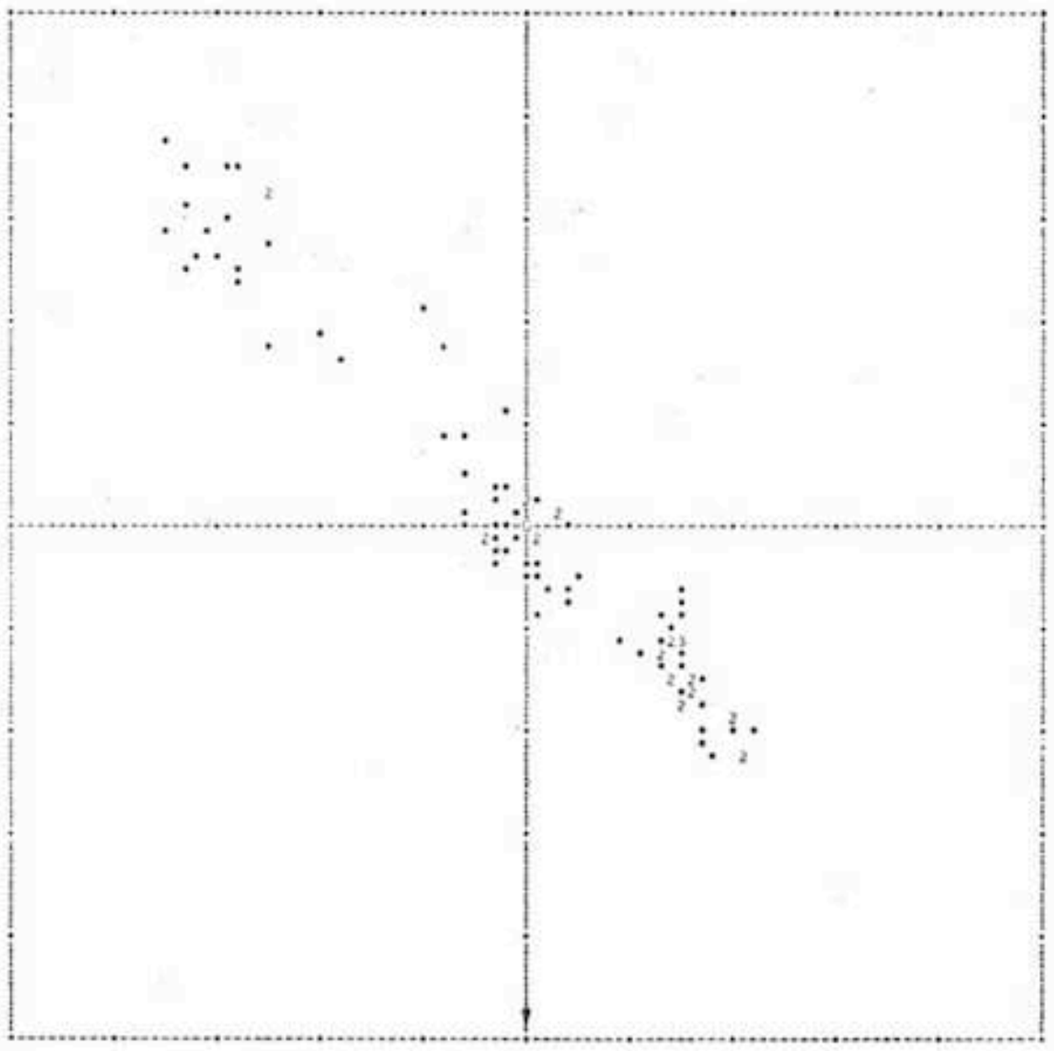

FIGURE 1. *First canonical correlation for 95 plankton samples in Pelorus Sound. The X-axis is a quadratic function of temperature and salinity and the Y-axis a linear function of the logarithms of counts of 11 species. The canonical correlation between the two functions is 0.9688.*

However, the equation itself is of very little interest. The coefficients of the ys provide no useful community diagnosis, and the coefficients of t and s tell us very little about the relationship to water properties except that it is non-linear (though the procedure for estimating the coefficients is, in effect, linear). The most useful feature is that it permits a fairly precise, though empirical, prediction of one side of the equation from the other (the canonical correlation is 0.97). This seems to provide a potentially useful tool for the ecologist. There are, for example some ecological situations, such as plankton sampling, where enumeration of organisms is tedious and time consuming, but many of the simpler environmental variables such as temperature and salinity can be (and often are) recorded by automatic instruments. Thus, provided a sufficient number of concurrent observations had been made to carry out a canonical correlation analysis, a great deal of the details of a plankton population could be filled in from more intensive temperature and salinity records of the habitat.

Conversely, on an intertidal mud-flat, many organisms are easily identified and counted in the field, but particle size and other sediment analyses are laborious. In mapping the sediments of a beach, it might then be economical to predict the sediment characteristics from the biota. This is merely an extension of regression sampling, a technique familiar in other fields but apparently little used by ecologists.

As this symposium has shown, there are many other multivariate techniques and many more problems to be solved. The above represents one approach. Another I am at present investigating is canonical analysis (a different technique from canonical correlation, see Seal 1964). By this method, the biota or their habitat are treated, not as continuous random variables, but as a series of discrete levels of organisation, separated by discontinuities. In some respects this is a more realistic model, because of the multiple cluster effect mentioned above, but it is too early to comment on the general utility of this technique.

REFERENCES

CASSIE, R. M. 1967. Mathematical models for the interpretation of inshore plankton communities. In *Estuaries* (Amer. Assoc. Advancement Sci.) pp. 509–14.

CASSIE, R. M. and MICHAEL, A. D. 1968. Fauna and sediments of an intertidal mud flat: a multivariate analysis. *J. Exper. Mar. Biol. Ecol.* 2: 1–23.

GOODALL, D. W. 1954. Objective methods in the classification of vegetation. III. An essay in the use of factor analysis. *Aust. J. Bot.* 2: 304–324.

ROHLF, F. J. 1968. Stereograms in numerical taxonomy. *Syst. Zool.* 17: 246–255.

SEAL, H. L. 1964. *Multivariate statistical analysis for biologists*. Methuen, London.

THURSTON, L. L. 1947. *Multiple-factor analysis*. Chicago.

WILLIAMS, W. T. and LAMBERT, J. M. 1959. Multivariate methods in plant ecology. I. Association analysis in plant communities. *J. Ecol.* 27: 83–101.