RR 91-02

GENERALISED BIPLOTS

John C. Gower

Department of Data Theory
University of Leiden

RR-91-02

GENERALISED BIPLOTS

John C. Gower

Department of Data Theory

University of Leiden

Generalised biplots

SUMMARY

A general but simple method of ordination is described that includes information on both samples and variables; new samples may be interpolated into the ordination. The method allows both continuous and categorical variables or mixtures of both types. It includes as special cases classical linear biplots, non-linear biplots and, for categorical variables, a new form of multiple correspondence analysis.

Some key words: Biplots; Non-linear biplots; Multiple correspondence analysis; Ordination; Multidimensional scaling; Graphical methods.

1. INTRODUCTION

The classical biplot of a multivariate sample (Gabriel, 1971) is an established tool for the initial exploration and display of both samples and variables, and relations between them. It is intimately related to principal components analysis (PCA) and, therefore, depends fundamentally on Pythagorean distance as a measure of inter-sample distance. The display of samples is precisely that of PCA; the variables are represented by vectors emanating from the centroid of the points representing the samples, that is from the point representing the sample mean. Gower and Harding (1988) showed how the biplot idea could be generalised for any inter-sample metric based on quantitative variables to give, as well as the usual ordination of the samples, representations of the variables as non-linear concurrent trajectories (see §2.2). They gave the details for Euclidean imbeddable metrics and ordination by classical scaling/principal coordinates analysis (PCO), where algebraic analysis is tractable. In principle, any form of ordination combined with any metric can be used, so long as there is some method for superimposing new samples in the space of the ordination and so long as all variables are quantitative.

This paper is concerned with a further generalisation, that removes the restriction to quantitative variables. Just as PCA and Pythagorean distance are the basis of the classical biplot for quantitative variables, multiple correspondence analysis (MCA) and chi-squared distance are the basis for the classical joint ordination of samples and categorical variables. MCA is concerned with the analysis of an n × p data-matrix X giving information on p categorical variates for each of n samples. Often X is represented as an indicator matrix G with n rows and $L = \sum_{k=1}^{p} l_k$ columns, where l_k is the number of categories that the kth variable may take. In the following, categories are referred to as category-levels to distinguish them from the categorical variables themselves. In the ith row of G each of the l_k columns associated with the kth variable is assigned to one of the category-levels and is scored unity if that level occurs in the ith sample, and zero otherwise. Thus the row-total of G for every sample is p, and the columns totals c_k (k=1,2,...,L) of G give the frequencies of each category-level of each variable. The kth variable cannot be represented by any form of trajectory but now must be represented by l_k discrete points.

The use of chi-squared distance is hard to justify when, as is usual, G is a sparse indicator matrix - see Greenacre (1990) for a discussion of the limitations of MCA. There is therefore a need to examine other approaches to the joint ordination of samples and variables described by multivariate categorical variables. The methodology to be developed will cater not only for categorical variables but also quantitative variables or mixtures of the two.

2. Incorporating Categorical Variables

2.1. Ordination of the Samples

In this section, first some comments are made on the ordination of samples described by categorical variables, and then it is shown how the non-linear biplot principle can be extended to plot points, in the space of the ordination, that represent the category-levels.

An obvious approach to the ordination of samples described by categorical variables is to treat the columns of G as L dichotomous variables and use one of the many similarity coefficients (see e.g. Sneath & Sokal 1973) to derive a matrix of similarities or dissimilarities, which can then be used as a basis of ordination by PCO, or any other form of metric or non-metric multidimensional scaling. This is indeed a viable possibility but there are some pitfalls.

The most simple similarity coefficient to use is the Jaccard coefficient but, because of the method of scoring G, matches get half the weight of mismatches. The simple matching coefficient gives l_k matches when the kth categorical variable matches and ($l_k - 2$) matches and two mismatches when the kth categorical variable mismatches. If one prefers equal weight for matches and mismatches, two possibilities are:

(i) Use the coefficient, which in the usual notation is written

$$s_{ij} = \frac{a}{a + \frac{1}{2}(b+c)}$$

which by giving half-weight to each mismatch, eliminates the mis-balance of the Jaccard coefficient which lacks the factor $\frac{1}{2}$. Gower & Legendre (1986) show that for this coefficient, $\sqrt{1 - s_{ij}}$ is a distance imbeddable in a Euclidean space.

(ii) Use the extended simple-matching coefficient for multi-level categorical variables (Gower, 1971). This contributes zero to the overall similarity of two samples if the kth categorical variable mismatches and unity if it matches, and may be evaluated from the entries of GG'.

The use of chi-squared distance in MCA generates a complicated weighting sytem which depends on the frequency of each categorical level in the sample (see §4.3); this is avoided by (ii), for which the algebraic details are developed in §4.2. The principle that I believe to be important is that weighting should be explicit and adopted for specific reasons and should not be implicit, obscure and dictated by intricate methodology; the generalised biplot gives control of weighting.

2.2. Incorporating Information on the Variables

If one of the approaches to ordination suggested in §2.1 is adopted, how then can information on the categorical variables be accomodated in the ordination diagram? A principal component analysis of G implicitly uses the simple matching coefficient, and the variables can be incorporated as in a classical biplot (Gabriel, 1971). However, we have seen that this coefficient gives differential weights to matches and mismatches and so may not be what is required.

For quantitative variables, Gower and Harding (1988) showed how non-linear biplots may be constructed for any dissimilarity coefficient and any ordination technique. Suppose X is an $n \times p$ data-matrix of quantitative variables, expressed as deviations from their means, and that a dissimilarity d_{ij} is defined between every pair (i,j) of rows of X. Using the notation $\{a_{ij}\}$, here and throughout this paper, to denote a matrix with typical element a_{ij}, an ordination of the dissimilarity matrix $D = \{d_{ij}\}$ consists in finding a set of coordinates Y $(n \times q)$ in q dimensions that generate a distance matrix $\Delta = \{\delta_{ij}\}$ that approximates, in some optimal sense, the dissimilarities D. This is achieved by defining and optimising a loss-function/goodness-of-fit criterion $S(\tau(D),\Delta)$. Here τ represents a transformation, either given or to be determined from the data, which may be applied to the original dissimilarities; often τ is the identity transformation, but polynomials, monotonic splines and montonic step functions are commonly used. An additional p-variable sample x may be superimposed on an existing ordination Y to give a point with coordinates y. This is done by optimising a criterion $T(\tau(d),\delta)$, where d is the vector of the dissimilarities between the new sample and the n original samples and δ is the corresponding vector of fitted values. The function τ is now always given, though it may have been determined empirically at the ordination stage. Superimposing an $(n + 1)$th point in this way differs from a simultaneous ordination of all $n + 1$ samples (i.e. the originals augmented by the new sample). For consistency, S and T should be the same function, though this is not absolutely necessary and might be relaxed, for example, if the use of different functions was found to be more efficient with little loss of accuracy. When S and T are the same function, the superimposition process is said to be coherent with the ordination process.

Non-linear biplots depend on superimposing artifcially constructed pseudo-samples x. The pseudo-sample x is defined to have zero values (representing the means) for all variables except for the kth variable which takes a value ρ. As ρ varies, the locus of the superimposed point y_ρ, corresponding to x, traces a trajectory assosciated with the kth variable (k = 1,...,p). The concatination of all p trajectories with the ordination Y defines the generalised biplot. The value ρ=0, corresponding to x being a sample of mean values, is common to all p trajectories, which are therefore concurrent.

The concept of a mean of categorical levels is not valid, so the role of zero values in the pseudo-variable must be reconsidered if the process is to be extended. To address this problem, consider the following bootstrap-like procedure. Construct a pseudo-sample by setting

$$x = (x_{j1},\ldots,x_{j(k-1)},\rho,x_{j(k+1)},\ldots,x_{jp})$$

and derive the superimposed point y_ρ on the ordination Y. Do this for j = 1,...,n and take the centroid of these n points. The trajectory for the kth variable is defined to be the locus of this centroid as ρ varies and the p trajectories together with Y form the generalised biplot. This may seem a somewhat arbitrary procedure, but it is shown in §3 that with ordination by PCO, and with a wide class of measures of dissimilarity that are in common use, it leads to the non-linear biplot for quantitative variables and, in §4, to a close relative to MCA for categorical variables. At the same time it includes many new special cases, the details of some of which are developed below, and admits mixtures of quantitative and categorical variables. That the procedure works well in this important special class of ordination procedures, suggests that it is worth considering with others and with non-Euclidean distances.

3. The Generalised Biplot in the Context of PCO

3.1. *The Basic Results*

The description of the generalised biplot given in §2 suffices for writing algorithms but gives no insight into the method and does not establish any properties that might be found useful for interpretation. To remedy this, algebraic analysis is desirable and this is most readily accomplished within the context of PCO; fortunately even with this limitation there is much of interest and there are many special cases of wide applicability.

Consider the nxp data-matrix X whose entries x_{ik} now may be either quantitative or categorical. Suppose also that squared-distance between the ith and jth samples is given by

$$d_{ij}^2 = \sum_{k=1}^{p} f(x_{ik}, x_{jk}) \tag{1}$$

i.e. each variable contributes independently to overall squared distance. The condition (1) is commonly satisfied by dissimilarity coefficients, and simplifies some of the following algebra, but it is not essential to the main argument. Let

$$D = \{-\tfrac{1}{2}d_{ij}^2\} \text{ and } D_k = \{-\tfrac{1}{2}f(x_{ik}, x_{jk})\} \text{ so that } D = \sum_{k=1}^{p} D_k,$$

and define the corresponding centred matrices

$$B = (I - N)D(I - N) \text{ and } B_k = (I - N)D_k(I - N) \text{ so that } B = \sum_{k=1}^{p} B_k$$

where N = ee'/n. Here, and throughout, e denotes a column-vector of units of length n, unless otherwise specified by a suffix. D_k and B_k give distance and centred matrices for the kth variable alone. When the n(n-1)/2 distances are imbeddable in Euclidean space the matrices B and B_k are positive semi-definite. Let the spectral decomposition of B be given by:

$$B = YY', \quad Y'Y = \Lambda,$$

where Λ is the diagonal matrix of eigenvalues of B. We have that Be = 0 and hence Y'e = 0; thus the rows of Y give the centred principal coordinates of D. It follows that the squared distances of each sample from the centroid of all samples is given by the elements of

$$\text{vec(diag B)} = \frac{e'De}{n^2}e - 2D\frac{e}{n}.$$

A new point, whose squared distances from the n points of the ordination are given in a vector f, is fixed in space and Gower (1968) showed that the coordinates y of its projection onto the space of the ordination are given by:

$$\begin{aligned} y &= -\tfrac{1}{2}(Y'Y)^{-1}Y'(f - \text{vec(diag B)}) \\ &= -\tfrac{1}{2}\Lambda^{-1}Y'(f + 2De/n). \end{aligned} \tag{2}$$

This formula gives the coherent (see §2.2) method for superimposing new points on a PCO ordination. In a PCO, only the first q columns of Y (ordered in decreasing order of the eigenvalues) will be used to give a q-dimensional approximation to D. In the following it is assumed that Y, and hence Λ, have q columns where $q \leq n-1$.

Consider the pseudo-sample proposed for generalised biplots, i.e. one that takes the same value as for the jth sample, except for the kth variable which takes the value ρ. The squared distance d_{ij}^{2*} of this sample from the ith sample is therefore given as

$$d_{ij}^{2*} = d_{ij}^2 - f(x_{ik}, x_{jk}) + f(x_{ik}, \rho).$$

These values may be formed into a matrix F whose jth column gives all the squared distances of the jth pseudo-sample from all the original samples. We have

$$F = -2D + 2D_k + f_k e' \qquad (3)$$

where $f_k = \{f(x_{ik}, \rho)\}$ is the column-vector giving the squared distances of the value ρ in the kth variable from the value of the kth variable in the ith sample. From the superimposition formula, the projections of all n pseudo-samples may be expressed in the present notation as:

$$Q = -\tfrac{1}{2}\Lambda^{-1}Y'(F + 2DN).$$

The jth row of Q gives the co-ordinates of the jth pseudo-sample projected onto the ordination space. Finally, the centroid of these n projections is:

$$\begin{aligned} Qe/n &= -\tfrac{1}{2}\Lambda^{-1}Y'(F + 2DN)e/n \\ &= -\tfrac{1}{2}\Lambda^{-1}Y'(f_k + 2D_k e/n), \text{ from (3).} \end{aligned} \qquad (4)$$

Note that $2D_k e/n = \frac{e'D_k e}{n^2} e$ − vec(diag B_k) which, apart from the ineffective constant term $(e'D_k e)/n^2$, gives the squared centroid distances for the kth variable alone. We shall use Q_k to label the point whose coordinates are given by (4).

Thus with PCO, equation (4) is the form taken by the method suggested in §2 as the basis for calculating the generalised biplot for the kth variable. As ρ changes, Q_k traces a trajectory when the kth variable is quantitative; for categorical variables ρ can take only l_k values and therefore Q_k defines l_k representative points, one for each category-level.

3.2. Interpolation

From the superimposition formula (2), the position in the ordination of a hypothetical sample $(\alpha_1,...,\alpha_p)$ is given by:

$$Z_\alpha = -\frac{1}{2}\Lambda^{-1}Y'(\sum_{k=1}^{p} f_k + 2De/n)$$

where $f_k = \{f(x_{ik}, \alpha_k)\}$ so that here ρ takes a different value, α_k, for each variable. Because of the additivity assumption $\sum_{k=1}^{p} D_k = D$, the above may be written

$$
\begin{aligned}
Z_\alpha &= -\frac{1}{2}\Lambda^{-1}Y'\sum_{k=1}^{p}(f_k + 2D_k e/n) \\
&= \sum_{k=1}^{p} Q_k = p\bar{Q}, \text{ say.} \qquad (5)
\end{aligned}
$$

Thus interpolation is simply a matter of finding the centroid $\bar{Q}$ of the points representing the values for the individual variables and extending this p times from the centroid G of the ordination Y; note that because $Y'e = 0$, the centroid G is at the origin. To avoid the inconvenience of extension, the points $R_k = pQ_k$ may be plotted rather than Q_k and this is done in the examples discussed in §5. This is purely a representational convenience and does not apply to any of the algebraic results given in the following.

Relative both to conventional rectilinear and to oblique coordinate axes, the interpolation rule given above gives the usual plot of a point $(\alpha_1,...,\alpha_p)$. The rule may therefore be regarded as a generalisation (i) for coordinate axes that are not necessarily linear and (ii) to accomodate categorical variables. For practical use, and precisely as is routine for linear axes, the non-linear trajectory axes for quantitative variables should be marked in unit steps of the original scales so that it is easy to locate α_k on the kth trajectory. Because of the non-linearity, equal steps on the original scales will not normally transform into equal steps on the non-linear axes; this can give useful information on the degree of local distortion in the approximation. Of course, for categorical variables the values of α_k are not numerical and are not represented by axes, but rather by the category-level points. However the centroid interpolation rule (5) remains valid. Indeed, as is shown in §3.4, the rule places the original samples precisely where they occur in the q-dimensional PCO ordination.

With conventional axes, each sample is uniquely identified by its coordinates. That samples are not uniquely defined with generalised biplots is a consequence of the q-dimensional approximation, rather than of the non-linear nature of the axes. Indeed, as is obvious from projection considerations, non-uniqueness is a property of classical linear biplots. With PCO in general, the non-uniqueness of projection manifests itself through the non-unique way in which the position of a fixed centroid can be generated by p points, one on each of p trajectories, or selected from representational point-sets, in q dimensions. When all variables are categorical, the finite number, L, of representative points induces a finite number, $\prod_{k=1}^{p} l_k$, of different centroids, which helps with interpretation. Similar remarks pertain to forms of ordination where the basis of approximation and superimposition is other than by projection.

3.3. Comparison with Non-linear Biplots.

In non-linear biplots (NLB) the value ρ=0 is common to all p trajectories which are therefore concurrent at a point O representing the superimposition corresponding to the mean of X. O coincides with the centroid G only for the classical linear biplot. The associated method for interpolation given by Gower and Harding (1988) differs from the generalised biplot (GB) interpolant of §3.2. Specifically, the NLB interpolant for $(\alpha_1,...,\alpha_p)$ is $\sum_{k=1}^{p} P_k - (p - 1)O$ relative to G, where P_k corresponds to α_k on the kth trajectory. This may be written as $p(\bar{P} - O) + O$, i.e. the centroid of the points P_k is extended relative to O (the original paper mentions an additional small correction representing the displacement of O from G but that is necessary only if the origin is restored to G). A more striking difference is that the GB trajectories derived from (4) for continuous variables are not concurrent (except in the classical case of Pythagorean distance, see §4.1). However, in both methods the vector f_k is the only term that varies with ρ, so the two sets of trajectories differ only in translation and will have the same shapes. Indeed, because both NLB and GB share the same ordination coordinates Y and the same superimposition formula (2), the interpolants for a sample $(\alpha_1,...,\alpha_p)$ described by continuous variables must be the same in both methods. The following gives a more algebric expression of this obvious geometrical result and shows how the NLB and GB trajectories may be transformed into one another.

When all variables are continuous, the trajectories arising from (4) can be translated to meet at O, so retaining the simple projection properties of the NLB approach. The interpolation method of §3.2 may be used to obtain the displacement of the trajectories given by the NLB from those given by (4) for the GB. To do this we need to know the position of the point O_k on the kth trajectory that corresponds to the mean of the kth variable; this is obtained from (4) by setting ρ=0 in f_k. The GB interpolant given by (4) for the NLB pseudo-sample (0,...,0,ρ,0,...,0) for the kth variable is then

$$P_k = O_1 + O_2 +...+O_{k-1} + Q_k + O_{k+1} +...+O_p$$

i.e.
$$P_k = p\bar{O} - O_k + Q_k,$$

where $\bar{O}$ is the centroid of $O_1, O_2,...,O_p$. However it follows from (5) that $p\bar{O}$ is the GB interpolant for the mean of X and hence coincides with O itself. Thus the interpolated position of the NLB pseudo-sample in the GB representation is simply

$$P_k = O - O_k + Q_k$$

relative to G. Of course P_k is, by definition, the point on the NLB which corresponds to Q_k on the GB; hence $O\text{-}O_k$ gives the displacement of the kth GB trajectory relative to the corresponding NLB trajectory. Because

$$\sum_{k=1}^{p} P_k = \sum_{k=1}^{p}(O - O_k) + \sum_{k=1}^{p} Q_k = (p-1)O + \sum_{k=1}^{p} Q_k,$$

it follows from the NLB interpolation result given at the beginning of this section, that the centroid-interpolants using either set of trajectories are identical. Thus with NLB trajectories, one extends $\bar{P}$ relative to O and with GB trajectories one extends $\bar{Q}$ relative to G. The two methods are precisely equivalent but the concurrency of the NLB trajectories is more convenient. This has shown that with quantitative variables, non-linear and generalised biplots are essentially the same thing, and simple translations can make them identical.

When categorical variables are introduced into GB, the method of §3.2 based on (4) may be used for the joint interpolation of categorical and quantitative variables. Although zero values are meaningless for categorical variables, the points O_k always exist for any associated continuous variables and their trajectories may be translated to concurrency at O as described above. The total interpolation relative to G may be written

$$p(\bar{P} - O) + p\bar{Q} + O.$$

Thus interpolation now requires a combination of the extension $p\bar{P}$ from O (continuous variables) with the extension $p\bar{Q}$ from G (categorical variables) and also involves a correction for the displacement of O from G; it may be more simple to sacrifice the convenience of having concurrent axes and use the direct method of §3.2.

This section has shown that the generalised biplot defined by (4) effectively subsumes the non-linear biplot for continuous variables of Gower and Harding (1988) and therefore both methods subsume the classical biplot for components analysis based on Pythagorean distance. However (4) is not confined to continuous variables but is also valid for categorical variables or, indeed, for mixtures of categorical and continuous variables.

3.4. Biplot Positions of the Sample Values

By letting ρ take in turn all the values x_{ik} of the kth variable we get n vectors f_i (i=1,...,n) which form the columns of the matrix $-2D_k$. Thus the projected coordinates for all the values are obtained by inserting into (4) to give

$$Z_k = -\Lambda^{-1}Y'(D_kN - D_k)$$

which may be written

$$Z_k = \Lambda^{-1}Y'(I - N)D_k(I - N)$$

or $$Z_k = \Lambda^{-1}Y'B_k. \qquad (6)$$

The columns of (6) give the n biplot coordinates associated with the n observed values of the kth variable. When the kth variate is categorical, B_k should be replaced by B_k^* formed from the l_k different columns of B_k and similarly Z_k should be replaced by Z_k^*. Otherwise (6) will unnecessarily repeat the same calculations. When the kth variable is quantitative, the n values of ρ that actually occur in the sample may be too few, or too unevenly scattered, to give an adequate description of the trajectory. Thus, equation (6) has disadvantages for computing trajectories but some useful algebraic results can be derived from it.

Adding the coordinates given by all k variables gives:

$$\begin{aligned}\sum_{k=1}^{p} Z_k &= \Lambda^{-1}Y'\sum_{k=1}^{p} B_k.\\ &= \Lambda^{-1}Y'B\\ &= Y'. \end{aligned} \qquad (7)$$

Formulae (6) and (7) are equivalent to the transition formulae of correspondence analysis. Formula (6) gives the coordinates for the variables in terms of the coordinates of the samples and (7) gives the inverse relationships. Incidently, (7) verifies that when (4) is used, the original samples project into their proper positions in the ordination. Indeed, this result follows direcly from (3) by setting F = – 2D, showing that the original samples are correctly interpolated even when the independence assumption (1) is not valid.

The mean of the n points given by (6) is $Z_k e/n = 0$. For categorical variables there are only l_k distinct points and if the ith level of the kth categorical variable occurs c_i times, then the weighted mean $\sum_{i=1}^{l_k} c_i z_{ik}/n$ is at the centroid of the ordination, where z_{ik} is obtained from the ith column of Z_k^* and contains the coordinates for the ith level of the kth variable. Note that this result is independent of how distance is defined for categorical variables, so long as the independence assumption (1) is valid.

3.5. Category and Sample Distances

The distances between pairs of sample points, pairs of category points and between sample and category points are evaluated explcitly in the following, and then some remarks are given on how to interpre these quantities. The squared distances between samples are given at the outset and, in terms of the elements of the centered, or inner-product, matrix $B = YY'$, may be expressed as $b_{ii} + b_{jj} - 2b_{ij} = d_{ij}^2$. This exemplifies the standard way for evaluating a matrix of squared distances from an inner-product matrix.

The equivalent of the inner-product matrix for two categorical variables h,k is $Z_h'Z_k$. From (6) $Z_h'Z_k = B_hB^-B_k$ where $B^- = Y\Lambda^{-2}Y' = (YY')^-$, the Moore-Penrose generalised inverse. The squared distances between pairs of different category levels of the same variable, and distances between category levels of different variables may be evaluated from these inner-product matrices, in the same way as for inter-sample distances.

Distances between samples and categories derive from the matrix $YZ_k = Y\Lambda^{-1}Y'B_k$. Now $Y\Lambda^{-1}Y' = I - MM'$ where the columns of M are the null-vectors of B. From the independence assumption, we have that if m is any vector, then $\sum_{k=1}^{p} m'B_km = m'Bm$. For Euclidean imbbeddable distances, all the matrices B_k are positive semi-definite, so that when m is a null vector of B it must also be a null vector of B_k ($k = 1,\ldots,p$). Hence $YZ_k = B_k$.

Gathering together all the results of the previous paragraphs leads us to consider the $n(p+1) \times n(p+1)$ symmetric inner-product block matrix Π, given by

$$\Pi = \begin{pmatrix} B & B_1 & B_2 & .. & B_p \\ B_1 & B_1B^-B_1 & B_1B^-B_2 & .. & B_1B^-B_p \\ B_2 & B_2B^-B_1 & B_2B^-B_2 & .. & B_2B^-B_p \\ .. & .. & .. & .. & .. \\ B_p & B_pB^-B_1 & B_pB^-B_2 & .. & B_pB^-B_p \end{pmatrix}$$

Because $BB^-B_k = B_k$, we may write

$$\Pi = P'B^-P$$

where $P = (B,B_1,B_2,\ldots,B_p)$, the first row/column of Π. Note that the sum of the elements of the ith row-block of Π is $2B_iB^-B$, which is $2B_i$, and hence that Π is in centred form.

The matrix Π contains all the information needed for a simultaneous ordination of the samples and variables of X. Although the size of Π makes its direct use impracticable, the spectral equation $\Pi U = U\Gamma$ may be solved, to give a simultaneous PCO,by noting that $B^- = Y\Lambda^{-2}Y'$ and setting $Q' = P'Y\Lambda^{-1}$. Then $\Pi = Q'Q$ and the spectral equation $QQ'V = V\Gamma$ satisfies $(Q'Q)Q'V=Q'V\Gamma$ showing that $U = Q'V$, after appropriate normalisation. The normalisation required is that $U'U = \Gamma$ which implies that $\Gamma = V'QQ'V = V'V\Gamma$, so we must normalise so that $V'V = I$. The advantage of this approach is that it requires the eigen-structure of the matrix

$$QQ' = \Lambda^{-1}Y'PP'Y\Lambda^{-1} = \Lambda^{-1}Y'(B^2 + B_1^2 + B_2^2 +\ldots+ B_p^2)Y\Lambda^{-1} \qquad (8)$$

of order q, which is usually very much smaller than the order n(p+1) of the matrix $\Pi = Q'Q$, even when $q = n - 1$.

Thus the spectral decomposition of QQ' gives a simultaneous PCO display of the np biplot points and the n samples, which is particularly attractive when all variables are categorical. Explicit expressions for all these coordinates may be obtained by substituting for P in $U = Q'V = P'Y\Lambda^{-1}V$. This gives the coordinates for the kth variable in this ordination as

$$\begin{aligned} U_k &= B_k Y \Lambda^{-1} V \\ &= Z_k'V, \text{ from (6).} \end{aligned}$$

Similarly the first n columns of P are those of the matrix B itself, giving the coordinates for the samples as

$$\begin{aligned} BY\Lambda^{-1}V &= (YY')Y\Lambda^{-1}V \\ &= YV. \end{aligned}$$

Thus the joint ordination merely rotates the previous biplot coordinates through a generalised angle given by the orthogonal matrix V. This shows that all the centroid properties obtained in §3.2 and §3.3 remain valid in the joint ordination. Any translation adjustments of trajectories for continuous variables, as discussed in §3.3, can be done after ordination.

With categorical variables, B_k repeats each level the number of times it occurs in G, thus giving a weighted analysis. This is easily remedied by replacing B_k in P by the $n \times l_k$ matrix B_k^* defined in §3.4, to give the $n \times (n+L)$ matrix $P^* = (B, B_1^*,...,B_p^*)$ and proceeding as before, replacing B_k^2 in (8) by $B_k^*B_k^{*'}$.

Defining $b_i^{(k)}$ for (k = 0,...,p) to be the ith column of B_k, which when k = 0 is defined to be a synonym for B, we have that the squared distance between the ith level of B_h and the jth level of B_k is given by the Mahalanobis-type metric

$$(b_i^{(h)} - b_j^{(k)})'B^-(b_i^{(h)} - b_j^{(k)}).$$

The matrix of these distances could be used with any form of multidimensional scaling to give a joint ordination of samples and variables.

Thus Π yields the above expressions for inter-sample, inter-variable, intra-variable and sample-variable distances. The interpretation of some of these distances requies care. Inter-sample distances are the usual approximations to d_{ij} as given by the ordination method, here PCO. The interpretation of inter-variable distances depends on whether distance is being measured between categorical or quantitative variables and whether between the same or different variables. A consequence of (6) is that when variables h and k (say) are such that $B_h = B_k$, then $Z_h = Z_k$ and the plotted points coincide. Thus if h and k are categorical variables with the same profiles in the samples, then their plotted points must coincide; if their profiles are similar the plots will show pairs of adjacent points, representing one level from each variable. Of course this can happen exactly only when h and k have the same number of levels (i.e. $l_h = l_k$). An aspect of the non-uniqueness of projection, and equivalently of the interpolation method discussed in §3.2, is that although common profiles imply adjacency, the converse is not necessarily true, so relationships inferred from the graphical plot should always be checked against the data. Even if the profiles match only for one category-level of each variable, then normally B_h^* and B_k^* will share a similar column so the corresponding category-level points will be adjacent; the details of how this happens depend on how the matrices B_k are calculated; specific examples are given in §4.2 and §4.3. When we are interested in distances between different category-levels of the same variable then h = k, and we are dealing with different columns of B_k^* which will pick out different parts of Y, so it is hard to make any general comments on the dispositions of the resulting pair of points although, as was shown in §3.4, the weighted mean of the points representing all l_k category-levels is at the centroid. Similar remarks apply to quantitative variables but there is now the possibility that two trajectories may overlap for whole or part of their course. Now it becomes important to distinguish between the observed sample values and other values that may have been interpolated to draw the trajectories. Even when the distances derived from Π are large, there may exist distances between interpolated points that are small, suggesting a correlational type of agreement between the two variables which can be deduced graphically.

Turning to the interpretation of sample-variable distance, it is clear from the form of Π that these distances are obtained from the centred matrices B_k for the individual variables, combined with the inter-sample and intra-variable information. The terms of Bk are intimately related to the pseudo-samples and their main interest is in the context of the interpolation results of §3.2. Because a sample is at the centroid of its associated category-level points, it might be hoped that some inverse relationship would hold, placing each category-level point at the centroid of those samples charecterised by that category-level. Unfortunately this seems not to be true but there is a sense in which something similar holds. Equation (6) shows that the coordinates for the variables are linear combinations of those for the samples. The precise form this takes depends on the form of B_k and special cases are elucidated in §4.2 and §4.3 where it is shown that, in these cases, a scaled form of the plotted points for the category-levels will be close to, (12), or coincident with, (20), the centroids of the samples with the corresponding category-levels. Little can be said in general but consider a set of samples which are identical in all their p category levels. Whatever measure of distance is used, these will necessarily have zero inter-sample distances and will be represented by coincident points in an ordination. If these category-levels occur in no other samples, this point must also represent the corresponding category-levels; if these category-levels do occur in other samples, then samples will be attracted towards the points representing their category levels.

With MCA, sample-variable relationships may be interpreted through the inner-product implicit in the singular-value decomposition, but it is unfortunate that the points plotted (see equations (13) and (14)) generate $p^{-1/2}U\Sigma^2V'C^{-1/2}$ rather than $U\Sigma V'$ itself. In general, this result is replaced by $YZ_k = B_k$, giving a simple, but not very helpful, cosine basis for interpretation.

4. Some Special Cases

4.1. Pythagorean Distance

For a continuous variable with Pythagorean distance we have $B_k = x_k x_k'$ where x_k gives the original (centered) sample-values of the kth variable. Then from (6)

$$Z_k = \Lambda^{-1}(Y'x_k)x_k'$$

which is of unit rank and hence the points represented by Z_k are collinear. Because $Z_k e = 0$, this line contains the centroid of the ordination; also from (4), the centroid of the ordination corresponds to the superimposed mean. This is in accordance with the classical linear biplot.

4.2. The Case of Categorical Variables with the Extended Matching Coefficient

Suppose d_{ij}^2 is given by the extended simple matching coefficient,

$$\text{i.e.} \quad \left.\begin{aligned} f(x_{ik},x_{jk}) &= 0 \qquad \text{if } x_{ik} = x_{jk} \\ &= 1 \qquad \text{if } x_{ik} \neq x_{jk} \end{aligned}\right\}.$$

Then, the full and kth-variable distance matrices are given by

$$D = -\tfrac{1}{2}(pee' - GG') \text{ and } D_k = -\tfrac{1}{2}(pee' - G_kG_k'),$$

where G_k comprises the l_k columns of G referring to the levels of the kth variable and C_k, which will be needed shortly, refers to the corresponding diagonal elements of C, so that $e'G_k = e_{l_k}C_k$. It follows from (6) that

$$\begin{aligned} Z_k &= -\tfrac{1}{2}\Lambda^{-1}Y'(I - N)(pee' - G_kG_k')(I - N) \\ &= \tfrac{1}{2}(1,2)\Lambda^{-1}Y'G_kG_k'(I - N) \\ &= -\tfrac{1}{2}(1,2)\Lambda^{-1}Y'\,G_k(C_ke_{l_k}e'/n - G_k')) \end{aligned} \qquad (9)$$

which is of size q×n but with only l_k distinct columns. Selecting these l_k columns gives

$$Z_k^* = -\tfrac{1}{2}\Lambda^{-1}Y'\,G_k(C_ke_{l_k}e_{l_k} - I_{l_k}) \qquad (10)$$

for the matrix giving the coordinates representing the kth categorical variable; the q × L matrix which gives the coordinates for all p variables is

$$Z = -\tfrac{1}{2}\Lambda^{-1}Y'\, G(CJ/n - I) \tag{11}$$

where J = diag $(e_{l_1}e'_{l_1}, e_{l_2}e'_{l_2}, \dots, e_{l_p}e'_{l_p})$, a block-diagonal matrix of order L.

Next we shall examine the detailed form given by (10) for the coordinates of the q × 1 vector z_1 representing level 1 of the kth variable. Equation (10) shows that the co-ordinates of level 1 depend essentially only on those rows of Y that refer to units having level 1 for the kth variable. Writing $\bar{y}_1$ for the mean of these rows of Y and similarly for $\bar{y}_2, \dots, \bar{y}_{l_k}$, (10) gives

$$z_1 = -\frac{1}{2}\Lambda^{-1}\left(\frac{1}{n}\sum_{i=1}^{l_k} c_i^2 \bar{y}_i - c_1 \bar{y}_1\right). \tag{12}$$

This result clearly generalises to other levels and other variables so that $\sum^{l_k} c_i z_i/n = 0$, showing that the weighted mean of the level co-ordinates for each variable is iad the centroid of the ordination, as was shown generally in §3.4. Note that for the unweighted mean

$$\sum_{i=1}^{l_k} z_i/l_k = -\frac{1}{2}\Lambda^{-1}\sum_{i=1}^{l_k} c_i^2 \bar{y}_i/n,$$

which may be regarded as a summary statistic for the kth categorical variate, from which each level co-ordinate, such as (12), deviates by the simple term $\frac{1}{2}c_i\Lambda^{-1}\,\bar{y}_i$ $(i=1,\dots,l_k)$. For equal weights $\sum_{i=1}^{l_k} z_i$ will be zero, and even for unequal weights, the weighted and unweighted means will often differ little, so that approximately

$$z_i \sim \tfrac{1}{2}c_i\Lambda^{-1}\bar{y}_i,$$

showing the category-level coordinates as a scaled form of the mean of sample coordinates, as referred to in §3.4.

In §3.4 it was shown that matching profiles imply matching trajectories. If only the first (say) level of h and k match then for the extended matching coefficient $f_h = f_k$ in (4) and adjacency then depends only on the relative values of the row-totals D_he and D_ke, or effectively on the centroid squared-distances for the two variables. The centroid distances may have similar values even when the other levels occur randomly giving, approximately, equal row-totals.

4.3. Multiple Correspondence Analysis

The multiple correspondence analysis of G is based on the singular value decomposition of

$$p^{-1/2}GC^{-1/2} = \frac{1}{n\sqrt{p}}ee'C^{1/2} + U\Sigma V'$$

to give sample coordinates $\quad Y = p^{-1/2}U\Sigma \quad$ (13)

and category coordinates $\quad Z = C^{-1/2}V\Sigma. \quad$ (14)

From (13) and (14) we have that

$$\begin{aligned} Z &= C^{-1/2}(V\Sigma U')U \\ &= p^{-1/2}C^{-1}G'U \\ &= C^{-1}G'Y\Sigma^{-1} \end{aligned}$$

or $\quad Z' = \Sigma^{-1}Y'GC^{-1}. \quad$ (15)

Similarly $\quad Y' = p^{-1}\Sigma^{-1}Z'G'. \quad$ (16)

Equations (15) and (16) are the transition formulae of correspondence analysis, slightly modified to accommodate the special conditions of MCA. The orthogonality conditions give e'Y = 0 and e'CZ = 0, so that Y gives centred principal coordinates of the points generating inter-sample chi-squared distances, and the weighted means of the category scores Z are also at the centroid.

The methodology developed in §3 may be used directly to operate on the matrix Y of MCA to derive the special form of category coordinates Z^+ obtained from (6) in accordance with the generalised biplot methodology. Accordingly, the formula for Z^+ is developed below, preparatory to comparing the Z of MCA (i.e. (15)) with Z^+. The chi-squared distance between two sample-units, one with the ith category-level and the other with the jth category-level of variable k, is given by

$$\left.\begin{aligned} d_{ij}^2 &= \frac{1}{p^2}\left(\frac{1}{c_i} + \frac{1}{c_j}\right) \qquad && \text{if } i \neq j \\ &= 0 && \text{if } i = j \end{aligned}\right\}.$$

Then, the full and kth-variable distance matrices are given by

$$D = -\tfrac{1}{2}p^{-2}(\Delta ee' + ee'\Delta - 2GC^{-1}G') \text{ and } D_k = -\tfrac{1}{2}p^{-2}(\Delta_k ee' + ee'\Delta_k - 2G_kC_k^{-1}G_k'),$$

where $\Delta = \text{diag}(GC^{-1}G')$ and $\Delta_k = \text{diag}(G_kC_k^{-1}G_k')$. It follows from (6) that

$$\begin{aligned} Z_k^+ &= p^{-2}\Lambda^{-1}Y'\,G_kC_k^{-1}(G_k' - C_ke\mathit{l}_ke'/n) \\ &= p^{-2}\Lambda^{-1}Y'\,G_kC_k^{-1}G_k' \end{aligned} \qquad (17)$$

which is of size q×n but with only l_k distinct columns. Selecting these l_k columns gives

$$Z_k^{+*} = p^{-2}\Lambda^{-1}Y'G_kC_k^{-1} \tag{18}$$

for the matrix giving the coordinates representing the kth categorical variable The q × L matrix giving the coordinates for all levels of all categorical variables obtained by (6) for chi-squared distance between samples is

$$Z^+ = p^{-2}\Lambda^{-1}Y'GC^{-1}. \tag{19}$$

Following very similar arguments to those given in §4.2, it follows from (6) that the coordinates of the first level of the kth variable are given by

$$z_1^+ = p^{-2}\Lambda^{-1}\bar{y}_1, \tag{20}$$

showing category-level coordinates as a scaled form of the mean of sample coordinates, as referred to in §3.4.

Comparing (15) and (19) we see that the only difference is in the scaling of the different dimensions. Indeed, because $\Lambda = Y'Y$ we have that $\Lambda^{-1} = p\Sigma^{-2}$, so that whereas the scaling associated with MCA is given by (15) as Σ^{-1}, that given by the generalised biplot of §4 is $p^{-1}\Sigma^{-2}$. Of course the distances generated by Z^+ are not chi-squared distances as are those given by Z, but this is of little moment for, as pointed out by Greenacre (1990), chi-squared distances generated between levels of possibly different categorical variables (the columns of G) have even less to recommend them than have the chi-squared distances between the rows of G.

Expressing Z^+ in terms of V gives:

$$Z^{+'} = p^{-1}C^{-1/2}V \tag{21}$$

which may be compared with Z given by (14), showing a close analogy with components analysis and classical biplot results where if, for centred X, we write $X = U\Sigma V$ then the scores for the units are given by $Y = U\Sigma$ (see (13)) and the variables are given by $Z = V$ (see (21)). Indeed a formal components analysis of $p^{-1/2}GC^{-1/2}$ (with non-centered G) may be used to derive the Y of MCA given by (13), and then either (15) used to obtain the MCA category scores or (21) used to obtain the alternative scores.

There are also strong similarities between the expressions for Z (15) and Z^+ (19) and that for Z (11) relating to the extended matching coefficient. A further comparison can be made between (20) and (12). However it must be remembered that the ordination Y in (11) differs from the MCA Y which is common to (15) and (19), so that direct comparison is invalid.

The argument given at the end of §4.2 that showed that for the extended matching coefficient, matches among only one level each of two categorical variables, could generate pairs of adjacent points, breaks down for MCA because of the differential weighting but it might be expected to remain approximately valid provided the weight-matrices C_h and C_k are not too disparate.

5. EXAMPLES

To illustrate the above methods, a small set of data given by Jongman, ter Braak and van Tongeren (1987) is used. This refers to 20 dune areas on the Dutch island of Terschelling, for each of which five environmental variables are available. The data is reproduced in Table 1. It is not intended to give here a detailed discussion of the data but merely to illustrate the similarities and differences of the various methods and to show the combination of categorical and continuous variables in a single ordination.

[Table 1, Figure 1, Figure 2 here]

Because moisture class M3 does not occur in any of the 20 samples listed, this category is absent in the subsequent analysis. Fig. 1 shows a two-dimensional MCA of the data of Table 1, excluding the quantitative variable for thickness of the A1 soil horizon. Fig. 2 shows the same data analysed using the extended matching coefficient (EMC) and the method of §3, as developed in §4.2. The two figures have much in common although, as is notorious for MCA, neither gives a good two dimensional fit (40% for MCA and 46% for EMC). The centroid interpolation results of §3 may be verified and it can be seen that MCA tends to draw towards the centroid those category-levels with the greatest weights (e.g. SF, U2, M1 and M5). The expected patterns are revealed in both plots. Thus nature management (NM) is associated with no fertiliser (C0) and high moisture content (M5) and these occur on farms 14, 15, 19 and 20, with other nature management farms with differing land use and moisture levels being more remote from the NM representative-point. Similarly, standard farms (SF) are associated with dry locations but not too dry (M2 rather than M1), high levels of fertiliser (C4) and mixed land use (U2) - these occur on farms 1,3,4 with near relatives on farms 12,13 and 16. Finally, hobby and biological farming are associated with low moisture (Ml), little fertiliser (C1 and C2) and grazing (U3). The full detail of Table 1 cannot be reproduced in the two dimensions of figures 1 and 2 and there are many intermediate samples not well integrated into the plots. Thus farm 16 with moisture-level M4 is plotted near the point for M2. Nevertheless the overall pictures given by Figs. 1 and 2 are remarkably accurate.

Not shown here is a MCA with the category-level points added by (4) in its special form (19). Of course the sample points are identical to those in Fig. 1. The remainder of the figure is so similar to Fig. 1 that it is not worth reproducing - the biggest difference is in the detail associated with BF and HF but is not of any significance. This similarity is not unexpected, as (15) and (19) differ only in scale and the singular values for the two dimensions, .81 and .75, are close enough to give nearly equal scaling on both axes whether squared or not.

[Figure 3 here]

In Fig. 3 the A1 horizon data, transformed to a log-scale and given unit range, is included and the variable Manure-class is treated as quantitative with unit range. Pythagorean distance is used in both cases. Thus in Fig. 3 these two variables are represented, as with classical biplots, by straight-lines through the centroid of the ordination. As recommended by Gower & Harding (1988) the plotted values are restricted to the range of these variables in the sample. The EMC was used for the three remaining qualitative variables and gave a similar plot to Fig. 2. Because the level of manuring is now treated quantitatively, it presents a more ordered sequence than previously. The deeper soil horizons tend to be associated with either nature management or standard farming and the shallower horizons with hobby/biological farming and dry soils; the line representing these quantities falls midway between the two groups of farms to accommodate the disparate groupings.

[Figure 4 here]

Fig. 4 is as for Fig. 3 but now the distance used for the two quantitative variables is the square-root of the L_1-norm. Gower & Harding (1988) explain how this gives a non-linear biplot for the variables which is piecewise-linear with "a corner" for every data-point. The two polygonal trajectories do not meet at the centroid, and indeed as was shown in §3.3, if there were more than two quantitative variables, then in general their corresponding trajectories would not be concurrent. The main difference from Fig. 3 is the way the polygonal line for manuring turns towards nature management for no manuring and towards standard farming for heavy manuring. For interpretation we need the projections of the mean for each quantitative variable, labelled as O_a for the Al-horizon and as O_c for manuring level. O_a is quite near O_c and the possibility of translating the two trajectories to O, the mean of O_a and O_c, to give the Gower-Harding non-linear biplot (see §3.3) is irrelevant in this case, especially as there are only two trajectories, so any inconvenience there may be in non-concurrency does not arise.

6. Conclusion

The above has developed a generalised biplot methodology which subsumes, as special cases, classical linear biplots and non-linear biplots for continuous variables It also allows categorical variables and mixtures of categorical and continuous variables; a different metric may be defined for every variable. The methodology has very great generality, embracing any ordination method. Details are given of the simplifications that occur when ordination is by PCO and variables contribute independently to an Euclidean imbeddable metric, thus giving a simple method for interpolating units in the ordination. A practicable general ordination method has been proposed which gives an optimal least-squares fit to n(p + 1) points, simultaneously representing samples and variables. In the special case when chi-squared distance is used for categorical variables, the ordination of the units must be that of MCA (apart, perhaps, form idiosyncrasies of scaling the axes); the generalised biplot ordination of the variables is shown to be very close indeed to that given by MCA itself. When the extended matching coefficient is used, a new form of ordination is derived which has advantages over MCA, while preserving some of the nice features of that method.

Even when the independence assumption is not satisfied, much of the basic methodology remains valid. Gower and Harding (1988) mentioned the possibility of using non-linear biplots with any form of metric scaling and with the so-called non-metric multidimensional scaling methods which employ monotonic, or other, mappings of the metric d_{ij}. This use has since been demonstrated in so-far unpublished work by Underhill interpolating into a monotonic step-function and by Heiser and Meulman interpolating into a smooth monotonic B-spline. These methods also form a special case of the generalised biplot. It is clear also that the method may be extended to more structured forms of sample (e.g. individual scaling, between/within groups canonical analyses, generalised Procrustes analysis). Every set of data may be regarded as a multidimensional configuration, embedded within which are the generalised biplot trajectory axes and points representing category-levels. It follows that any transformations done in the course of subsequent data-analysis may be made to operate on the augmented configuration, so carrying over into the final analyses, biplot information which may be used to aid interpretation. Thus the simple idea underlying the generalised biplot both unifies and greatly extends exisiting methodology.

ACKNOWLEDGEMENTS

I thank Kristel Luijtens for programming in Genstat the calculations needed for the examples. The first version of this paper was written while the author was employed by The Lawes Agricultural Trust, Rothamsted Experimental Station.

REFERENCES

GABRIEL, K.R. (1971) The biplot-graphic display of matrices with applications to principal components analysis. *Biometrika* 58, 453-67.

GOWER, J. C. (1968) Adding a point to vector diagrams in multivariate analysis. *Biometrika* 55, 582-5.

GOWER, J. C. (1971) A general coefficient of similarity and some of its properties. *Biometrics* 27, 857-71.

GOWER, J. C. & HARDING, S. (1988) Non-linear biplots. *Biometrika* 73, 445-55.

GOWER, J. C. & LEGENDRE, P. (1986) Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification* 3, 5-48.

GREENACRE, M. J. (1988) Correspondence analysis of multivariate categorical data by weighted least squares. *Biometrika* 75, 457-65.

GREENACRE, M. J. (1990) The limitations of multiple correspondence analysis. *Computational Statistics Quarterly* 3, 249-56.

HEALY, M. J. R., GOLDSTEIN, H. (1976) An approach to the scaling of categorised attributes. *Biometrika* 63, 219-229.

JONGMAN, R. H. G., TER BRAAK, C. J. F. & VAN TONGEREN, O. F. R. (1987) *Data analysis in community and landscape ecology*. Wageningen: Pudoc (Centre for Agricultural Publishing and Documentation).

SNEATH, P. M. A. & SOKAL, R. R. (1973) *Numerical Taxonomy*. San Francisco: W. H. Freeeman & Company.

Sample number	Al horizon (cms)	Moisture class	Grassland management type	Grassland use	Manure class
1	2.8	1	SF	2	4
2	3.5	1	BF	2	2
3	4.3	2	SF	2	4
4	4.2	2	SF	2	4
5	6.3	1	HF	1	2
6	4.3	1	HF	2	2
7	2.8	1	HF	3	3
8	4.2	5	HF	3	3
9	3.7	4	HF	1	1
10	3.3	2	BF	1	1
11	3.5	1	BF	3	1
12	5.8	4	SF	2	2
13	6.0	5	SF	2	3
14	9.3	5	NM	3	0
15	11.5	5	NM	2	0
16	5.7	5	SF	3	3
17	4.0	2	NM	1	0
18	4.6	1	NM	1	0
19	3.7	5	NM	1	0
20	3.5	5	NM	1	0

Table 1: Environmental information for 20 Dutch dune sites (reproduced with permission from Jongman, ter Braak and van Tongeren, 1987)

Key; Grassland management (standard farming SF, biological farming BF, hobby farming HF, nature conservation management NM), Grassland use (hay production 1, intermediate 2, grazing 3), Moisture class and Manure class are ordinal variables.

Figures 1 - 4. These are all ordinations of Table 1.
Figure 1 is a Multiple Correspondence Analysis, Figure 2 is an analysis based on the Extended Matching Coefficient, Figure 3 introduces continuous variables with Pythagorean distance and Figure 4 continuous variables with the square-root of the L_1-norm.

Key
▲ Grassland management: SF, BF, HF, NM (see key to Table 1)
■ Grassland use: U1 (hay production), U2 (intermediate), U3 (grazing)
● Moisture class: M1, M2, M4, M5
O Manure class: C0, C1, C2, C3, C4 (levels of manuring)
× Al horizon in log-centimetres
✱ G (Centroid), O_a (mean of Al horizon), O_c (mean of manure levels)

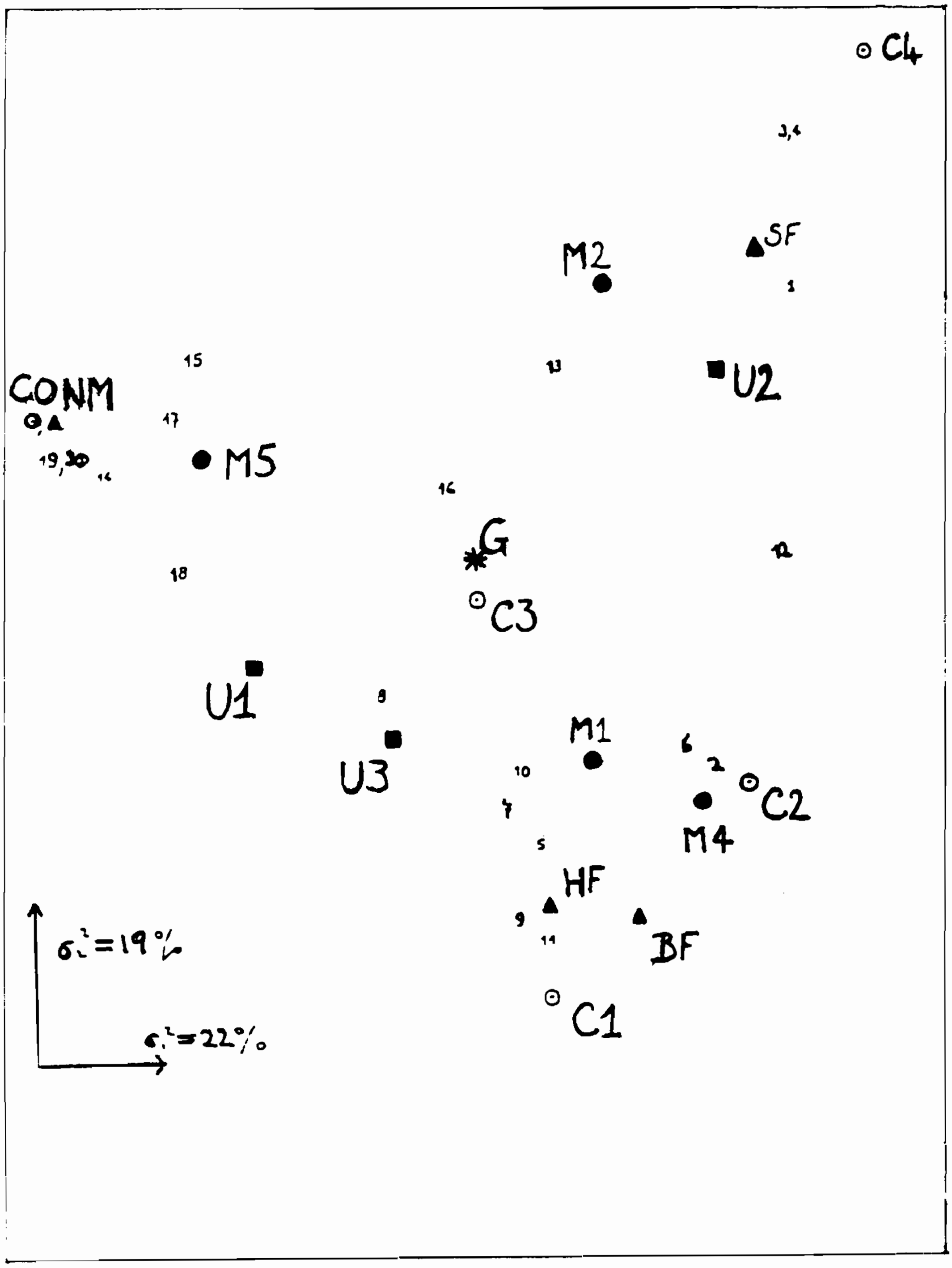

C4
3,4
SF
M2
1
13
U2
15
CONM
17
19,20
14
M5
16
G
12
18
C3
U1
8
U3
M1
6
2
10
C2
7
M4
5
HF
9
BF
11
C1
$\sigma_2^2 = 19\%$
$\sigma_1^2 = 22\%$

Figure 1

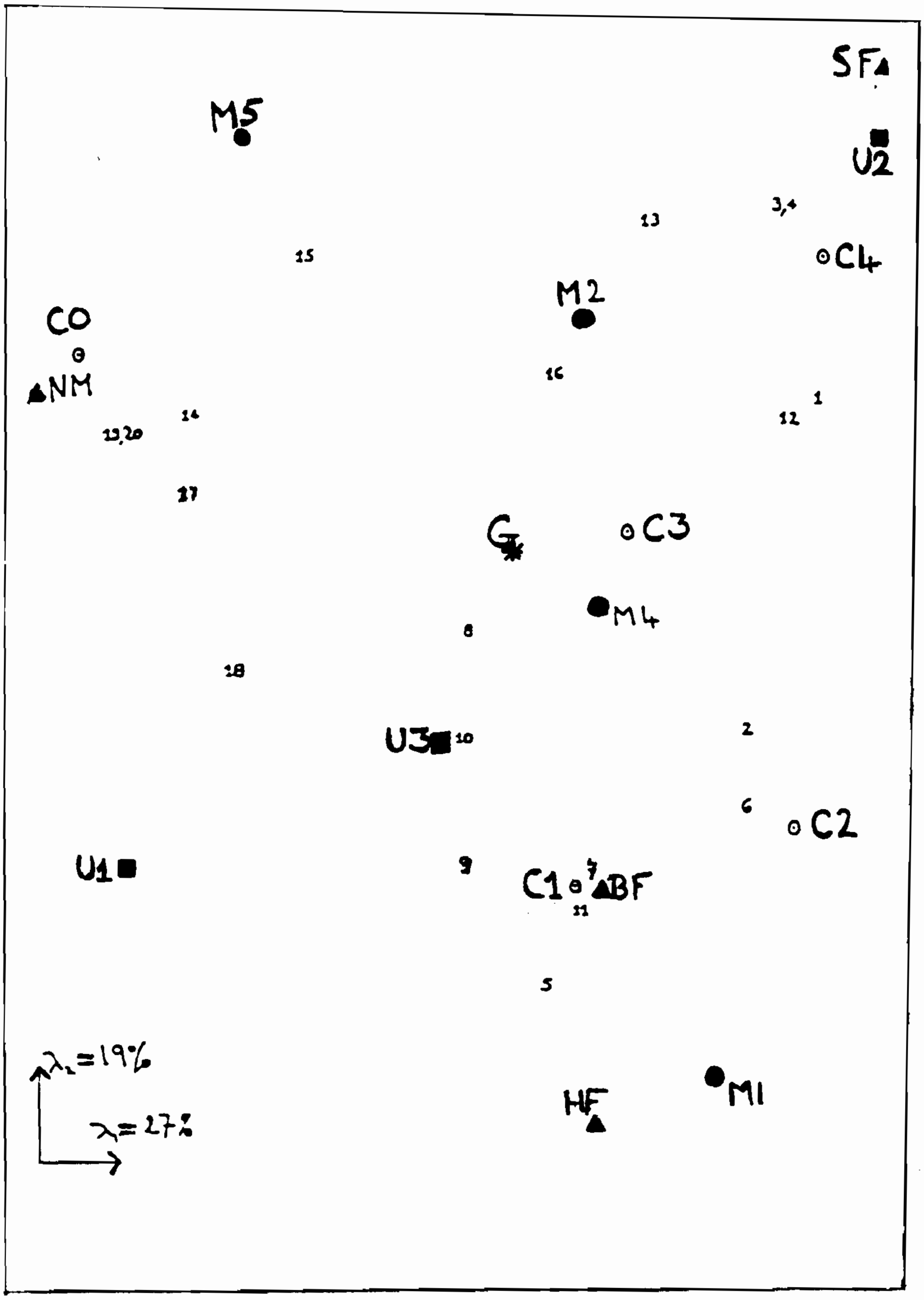

SF
M5
U2
3,4
13
15
C4
M2
CO
NM
16
1
12
14
19,20
17
G
C3
M4
8
18
U3
10
2
6
C2
U1
9
C1
7
BF
11
5
$\lambda_2=19\%$
$\lambda_1=27\%$
HF
MI

Figure 2

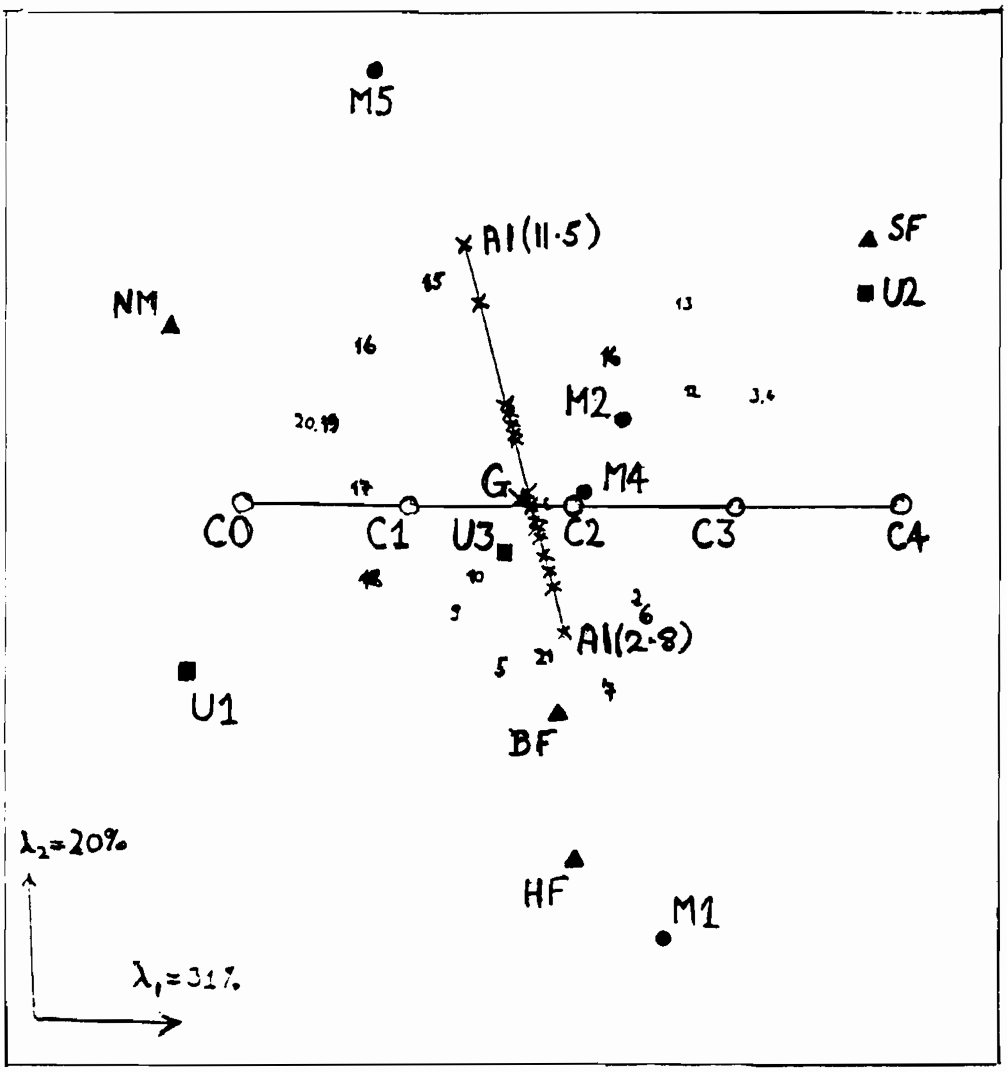
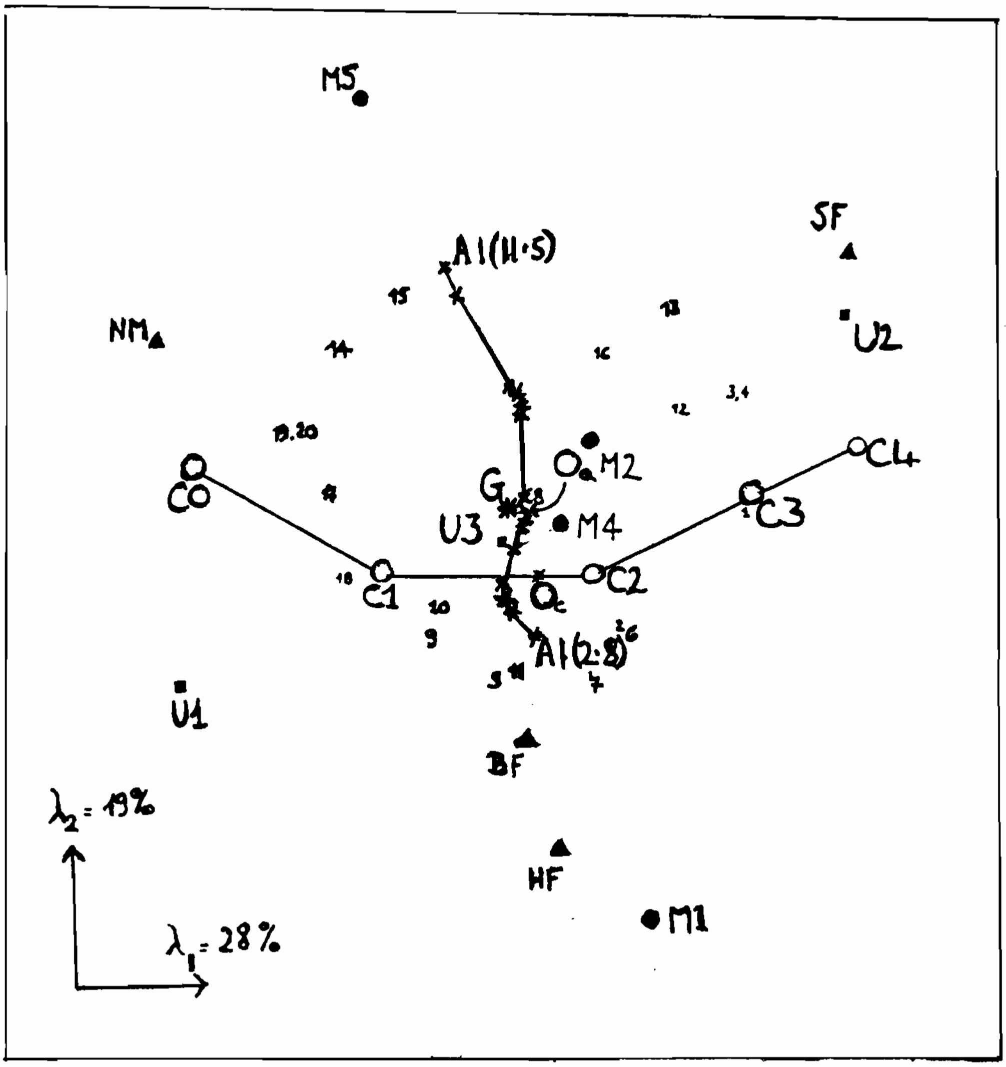

M5
A1(11·5)
SF
U2
15
13
NM
16
14
M2
12
3,4
20,19
G
M4
17
C0
C1
U3
C2
C3
C4
18
10
2
6
9
A1(2·8)
5
21
7
U1
BF
λ_2=20%
HF
M1
λ_1=31%

Figure 3



M5
SF
A1(11·5)
15
13
NM
U2
14
16
3,4
12
19.20
CO
M2
C4
G
C3
U3
M4
C1
C2
18
20
A1(2·8)
2
6
9
5
4
U1
BF
$\lambda_2 = 19\%$
HF
M1
$\lambda_1 = 28\%$

Figure 4