Developments in Applied Statistics
Anuška Ferligoj and Andrej Mrvar (Editors),
Metodološki zvezki, 19, Ljubljana: FDV, 2003

Unified Biplot Geometry

John C. Gower[1]

Abstract

The fundamental geometry is outlined that underlies all biplots of a data-matrix **X** of *n* cases and *p* variables. Cases are represented by *n* points and variables by a *reference system*. The reference system for quantitative variables may be orthogonal Cartesian axes, other linear axes or nonlinear trajectories. The reference system for categorical variables is a set of *category-level-points* (CLPs) one for each category-level; CLPs for ordered categories are collinear. Axes are labelled by a set of graduated numerical markers; CLPs are labelled by the names of their category levels. The point representing a case is nearer the markers that give the values of its variables, than to any other markers. This high dimensional representation is approximated in few (often two) dimensions in such a way that the approximated reference system gives optimal approximations to the values of **X**. Furthermore, new cases may be interpolated into the approximation space. Special cases within this general framework are illustrated by several examples of biplots.

1 Introduction

Most of the material presented here is based on the book by Gower and Hand (1996), where algebraic details may be found; some more recent material is briefly reviewed at the end. These notes give some examples and an overview of the geometrical underpinning of biplots.

Biplots are the multivariate analogue of scatter plots. Multidimensional Scaling (MDS) is used to approximate the multivariate distribution of a sample in a few dimensions, typically two, and superimposed on this display are representations of the variables on which the samples are measured. In this way the relationships between the individual sample points can be easily seen and related to values of the measurements. Like scatter plots, biplots are useful for giving a graphical description of the data, for detecting patterns which possibly lead to more formal analyses, and for displaying results found by more formal

[1] The Open University, Department of Statisitics, Milton Keynes MK7 6AA, U.K.

methods of analysis. The *bi* in biplots denotes that both samples and measured variables are represented, not that biplots are necessarily two-dimensional, though they usually are.

Approximations to the relationships between n samples can be achieved by the various methods of MDS but, except in the special cases of principal components analysis (PCA), multiple correspondence analysis (MCA), and canonical variate analysis (CVA), until recently methods for including information on the variables have been little developed. In recent years, the theory of biplots has been considerably extended, and can now be presented and extended in a unified manner that includes biplots for PCA, CVA and MCA as special cases, as well as some newer and less well-known methods; there is plenty of room for further development.

The concept of inter-sample distance (d_{ij}) is central to all methods of MDS, and this is one of two components of the unifying concepts that underpin all that follows. In MDS the samples are represented by n points in a low-dimensional space that *generate* approximations δ_{ij} to the given distances d_{ij}. In MDS, different methods may be used for calculating the distances- Pythagorean distance for PCA, chi-square distance for MCA, Mahalanobis distance for CVA, any Euclidean embeddable distance for principal coordinates analysis (PCO), and many others - and different methods of MDS use different kinds of approximation. Because Euclidean displays, with which most research workers are familiar, form the overwhelming majority of published material, attention will be confined mostly to Euclidean distances and totally to Euclidean displays; this constraint could be relaxed, as it is occasionally in MDS. The other component of unification that is to be discussed is strongly based on the familiar notion of coordinate axes. Figure 1 shows the conventional Cartesian coordinate system for two quantitative variables and emphasises the distinction between (i) *interpolation:* i.e. assigning a point with given values of the variables, which is done as a vector-sum, and (ii) *prediction:* i.e. associating values of the variables with a given point, which is done by orthogonal projection. As is obvious from the figure, these two operations are consistent. Consistency holds only in exact representations and breaks down in MDS approximations in fewer dimensions than there are variables. As we shall see, these require separate coordinate representations for interpolation and prediction.

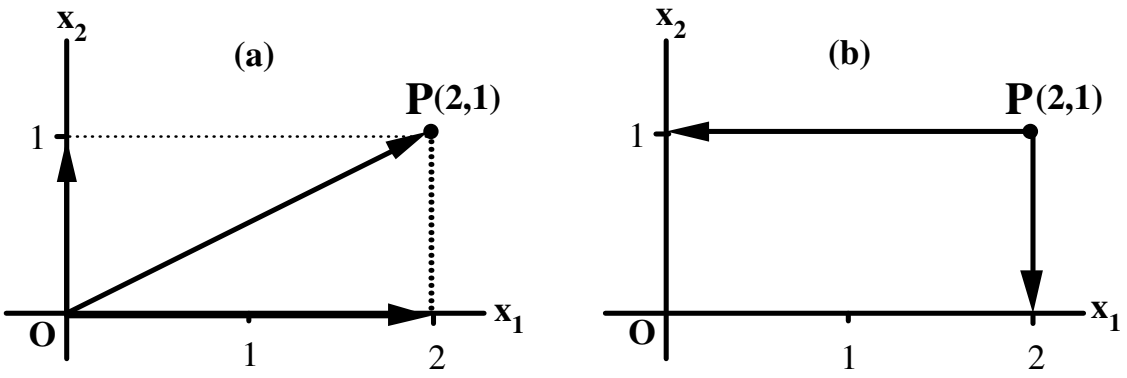

Figure 1: Coordinate axes (two dimensions) (a) illustrates interpolation and (b) prediction.

Mention of PCA and MCA indicates an interest in both continuous and categorical variables. While a continuous variable will be represented by a biplot axis which is a continuous curve (not necessarily linear) labelled by values of the variable, a categorical variable will be represented by a biplot "axis" which is a simplex of points, the category level points (CLPs), labelled by category names. Both types of representation may occur in the same plot and the set of such generalised axes is termed a *reference system*. Much multivariate analysis is concerned with canonical axes of one kind or another. The diagrams shown will not contain the usual rectangular canonical axes (indeed; there are more "axes" than displayed dimensions) but the reference system of biplot axes (including CLPs) serves a similar purpose. In the unified approach, interpretation of biplots is firmly based on coordinate axes representing the original variables, at the same time extending the notion to include representations of categorical variables.

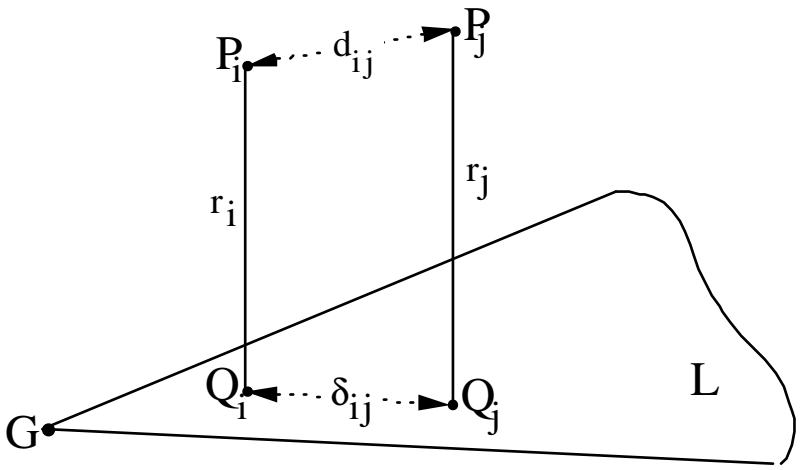

Figure 2: G is the centroid of all n points and lies in L. Q_i and Q_j are the orthogonal projections of P_i and P_j onto L and r_i and r_j are the residuals from L.

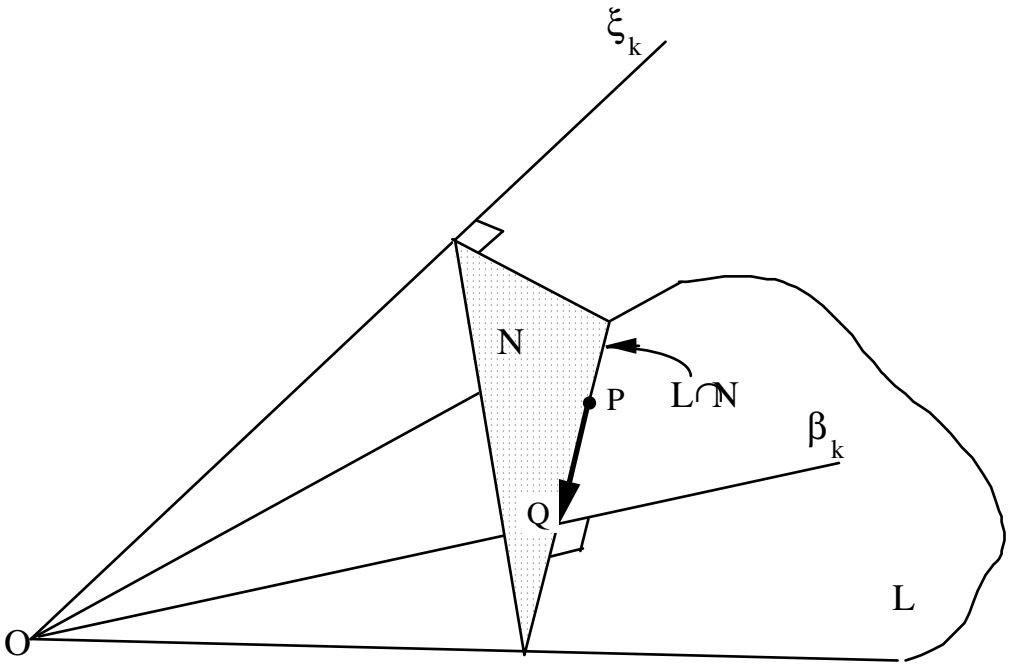

Figure 3: Representation of a coordinate axis ξ_k by a biplot axis β_k in the space L (plane) of approximation.

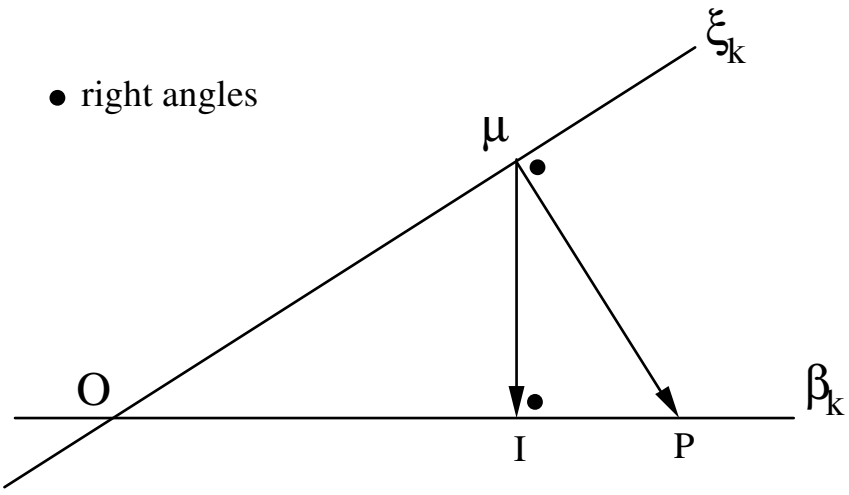

Figure 4: Relationship between the markers for prediction and interpolation. ξ_k is one of the axes in R and β_k is the corresponding biplot axis in L. The marker μ projects to I for interpolation and back-projects to P for prediction.

2 Principal components biplots (classical biplots)

In PCA observed distances d_{ij} are replaced by approximate distances δ_{ij} using orthogonal projection, as shown in Figure 2. The approximation in L is obtained by minimising the sum of squares of the residuals and L is a subspace of the exact representation. In PCA d_{ij} is the Pythagorean distance between P_i and P_j.

Figure 3 shows what happens to a coordinate axis ξ_k. This too is projected onto L to become the *biplot axis* β_k. The plane N contains all points with coordinate x_k on the axis ξ_k. In particular, the intersection L∩N contains all points in L that predict the value x_k.The biplot axis β_k is orthogonal to the intersection and hence the projection of the point P onto β_k gives the correct prediction, provided a suitable scale is marked on β_k. All the information necessary for prediction is contained in the approximation space L. Things are different for interpolation where β_k remains in the same direction but the marker for x_k is now obtained by orthogonal projection from ξ_k onto β_k. Figure 4 shows the relationship between the markers for interpolation and prediction.

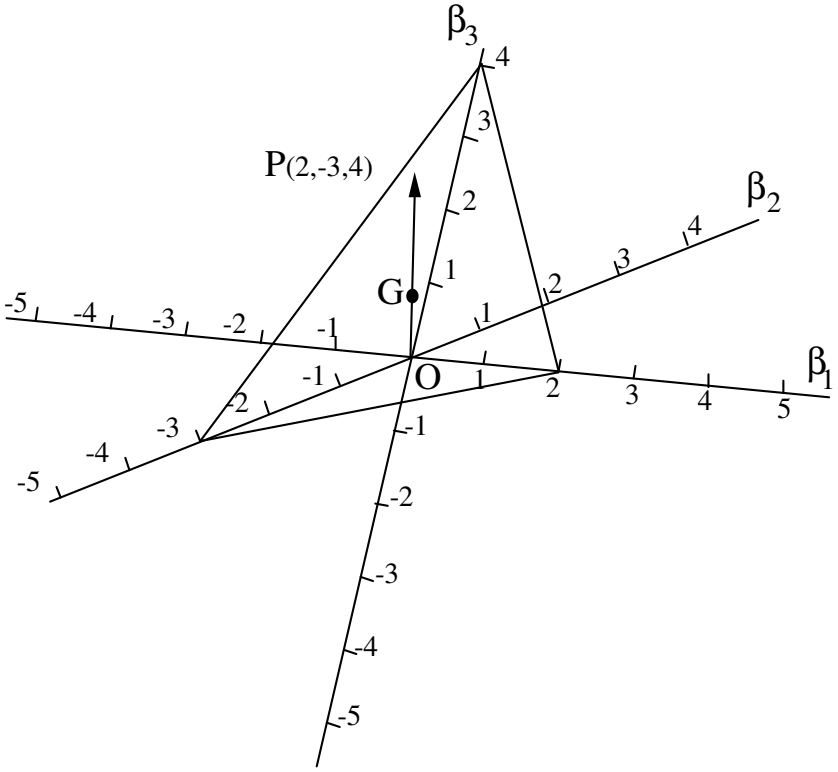

Figure 5: Interpolation, using the vector-sum method, of the point (2,-3,4). G is the centroid of the markers 2,-3 and 4 on the biplot axes β_1,β_2, and β_3 (respectively). The interpolated point P is at three times OG.

Figure 5 shows how interpolation may be obtained as a vector-sum. Rather than "completing parallelograms", it is simpler to find the centroid of the relevant markers and extend p (the number of variables) times from the origin. Here $p = 3$ so we have a reference system for three variables in a two dimensional approximation; the biplot axes are necessarily non-orthogonal.

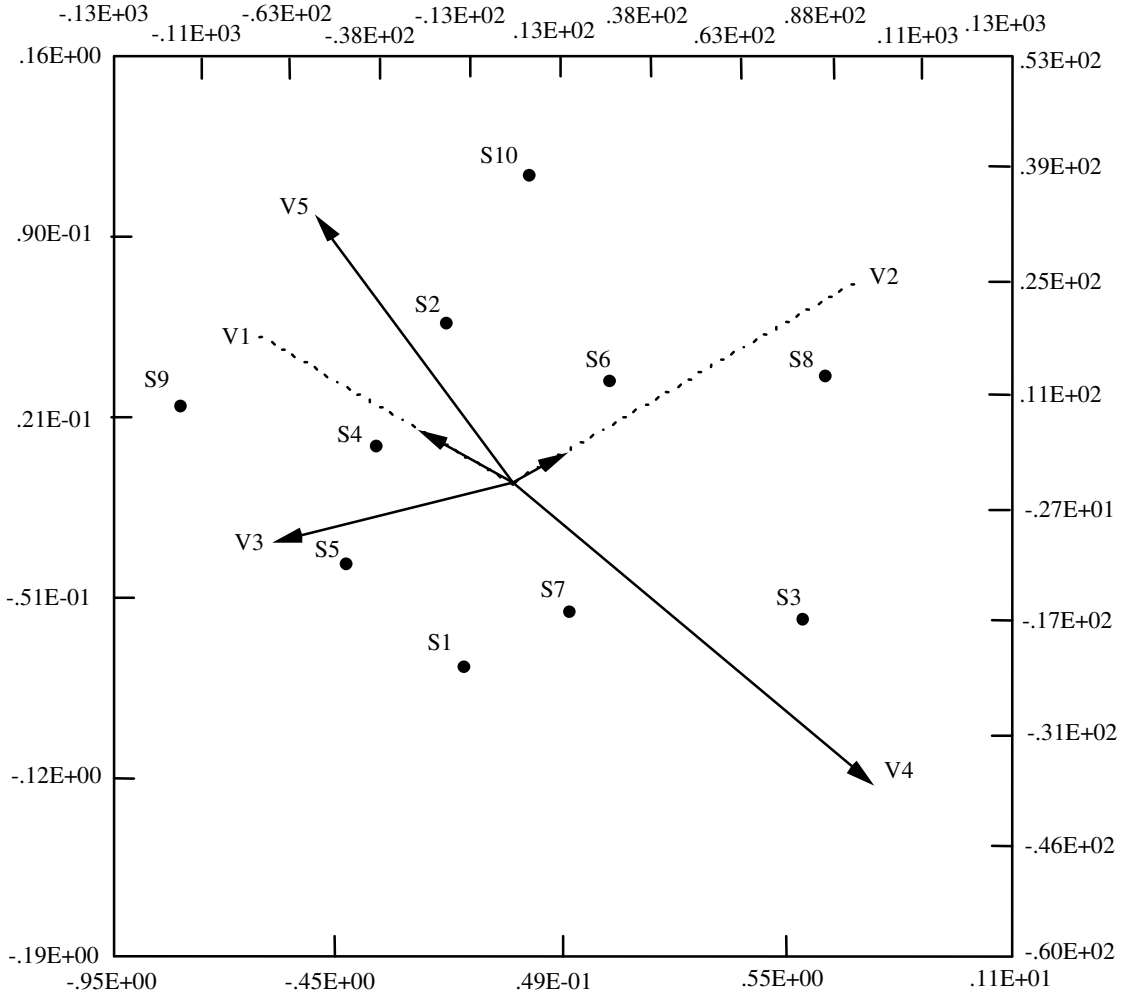

Figure 6: A fictitious example illustrating some of the faults often to be found in biplot representations.

Figure 6 shows some of the faults often found in published biplot diagrams. Unequal scaling of the axes, use of E-formats, ugly scale divisions, no scales on biplot axes, one-standard error lengths of vectors, scales given unnecessarily for canonical axes and no scales for the original variables, separate canonical scales for samples and for variables.

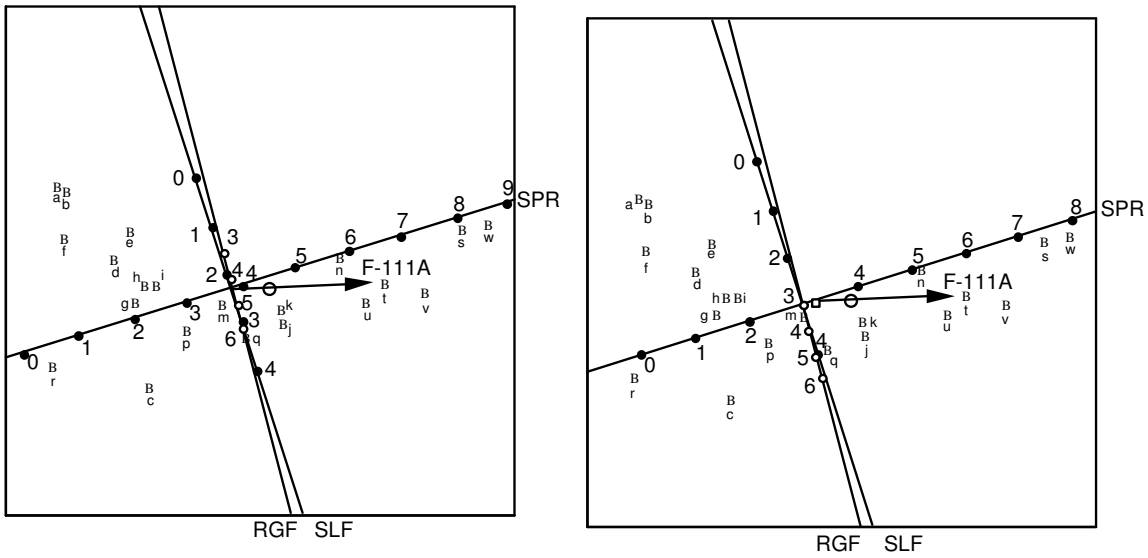

Figure 7: Interpolative PCA biplots without and with a translation so that the value "3' coincides at the origin. The small open square represents the centroid from which vector-sum interpolation is appropriate in both cases.

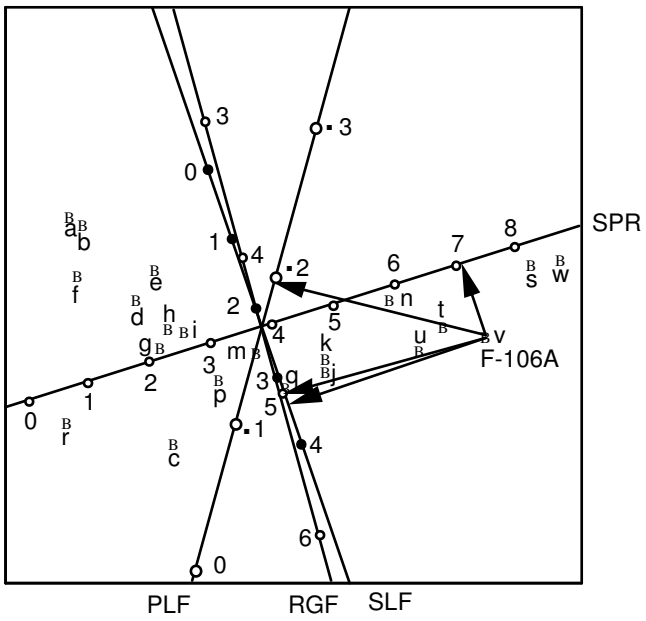

Figure 8: Predictive PCA biplots for the same data as in Figure 7.

Figure 7 shows an interpolative classical biplot for data on four variables relating to fighter aircraft. The right-hand figure differs from the left-hand only in that the axes have been translated so that a whole number (here, 3) occurs at the origin. This kind of simplification is possible for interpolation so long as the vector-sum of the translations is null; it is not permissible in predictive biplots. The fourth variable PLF has a very small range and contributes only marginally to the display so is not exhibited. Figure 8 shows the corresponding predictive biplot

which involves all four variables; note that the axes have the same directions as in Figure 7 but different spacing for the scalings of the markers.

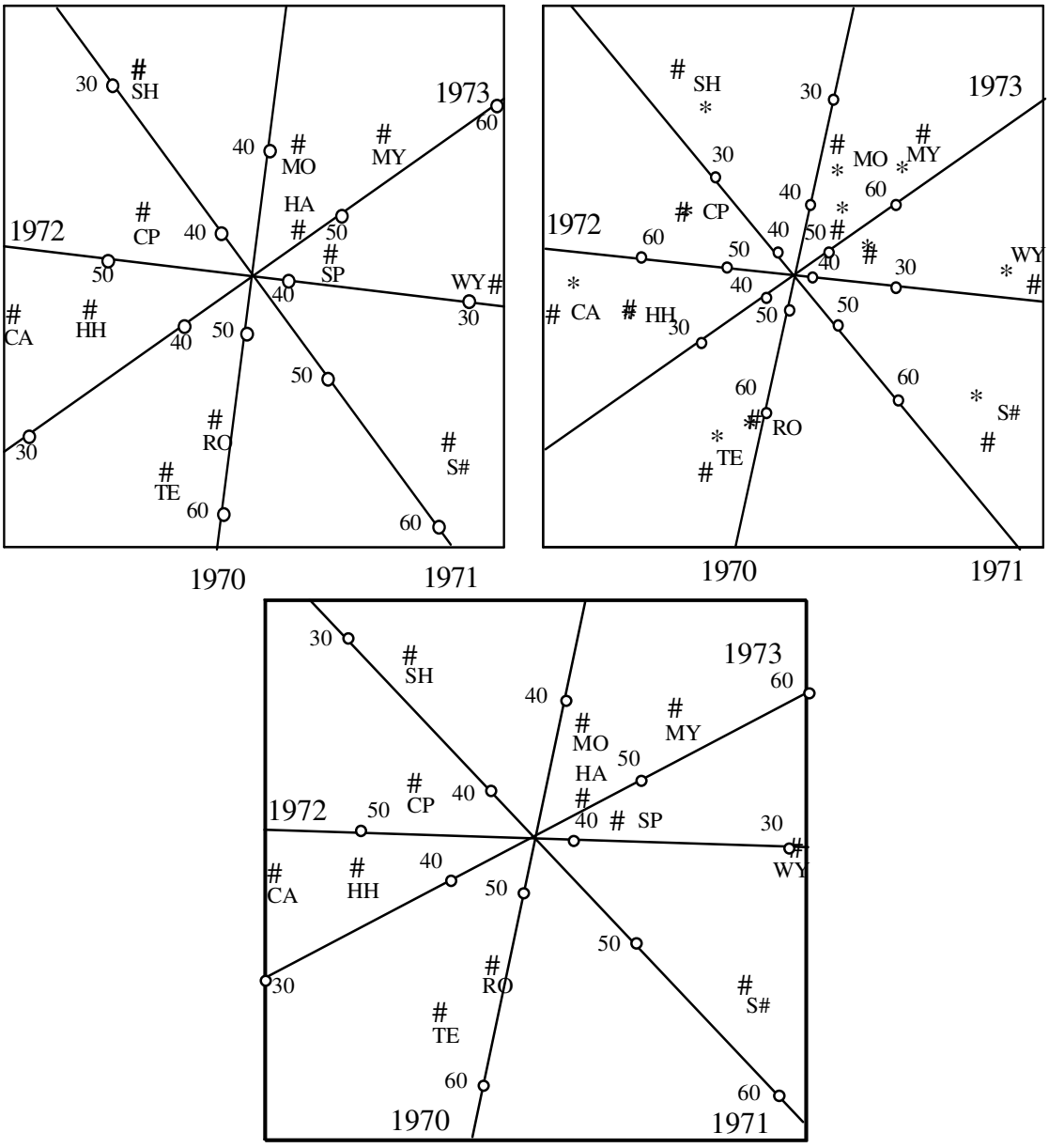

Figure 9: Biplots for least-squares metric multidimensional scaling (minimisation of stress). Predictive biplot using Procrustean embedding, Approximate interpolative biplot using minimum error projection Procrustean embedding and a predictive biplot using regression embedding.

3 Other linear biplots

Now we assume that d_{ij} remains defined by Pythagorean distance but is approximated by other methods of metric MDS, in particular by minimising the Stress and Sstress criteria, $\sum_{i<j=1}^{p}(d_{ij}-\delta_{ij})^2$ and $\sum_{i<j=1}^{p}(d_{ij}^2-\delta_{ij}^2)^2$, respectively.

The example shown in Figure 9 relates to wheat yields at twelve centres in four years. Interpretation is very similar to that of the classical biplots. The special interest is that L is not now a subspace of the full exact representation and must somehow be embedded in the full space. Three ways of doing this are: (i) By orthogonal Procrustes fitting, (ii) By multiple regression and (iii) By projection. If **X** is the data-matrix and **Z** the coordinates in L that generate the δ_{ij}, then these may be represented by minimising $\|\mathbf{X} - \mathbf{ZA}\|$ where **A** is, respectively an orthogonal matrix, a general matrix and a projection matrix. All three methods give the same result when **Z** is obtained by PCA. Once L is embedded by one of these methods, the geometry of Figure 3 remains valid and prediction proceeds as with PCA, merely using a different sub-space L. The first (Procrustean) and last (Regression) biplots are both predictive and can be seen to be very similar. Because of technical difficulties, geometric (as opposed to algebraic) interpolation with general methods of metric scaling seem to be beyond reach. An approximation can be found by choosing L to be the subspace for which the projections of **X** onto L best match **Z**; this is the minimum error projection Procrustes method and is shown in the middle plot of Figure 9. The asterisked points are obtained by projection and, for comparison, the hash-symbols reproduce the Procrustean embedding of the first plot. Of the three methods considered, Regression must minimise the prediction error but it is not clear whether this optimal property necessarily extends to other choices of distance, especially to non-metric MDS.

4 Categorical variables: multiple correspondence analysis and related methods

The data-matrix **X** now consists of categorical variables but the algebra may continue to be presented as a PCA of quantitative data. Each category-level is represented by a single variable and in **X** is replaced by the square root of its frequency in the sample. d_{ij} is now chi-squared distance. **L** is a diagonal matrix giving the frequencies of all the categories. The rows of $\mathbf{L}^{-1/2}$ give the coordinates of the CLPs and may be projected like linear biplot axes but, rather than a continuum, they give a single point for each level.

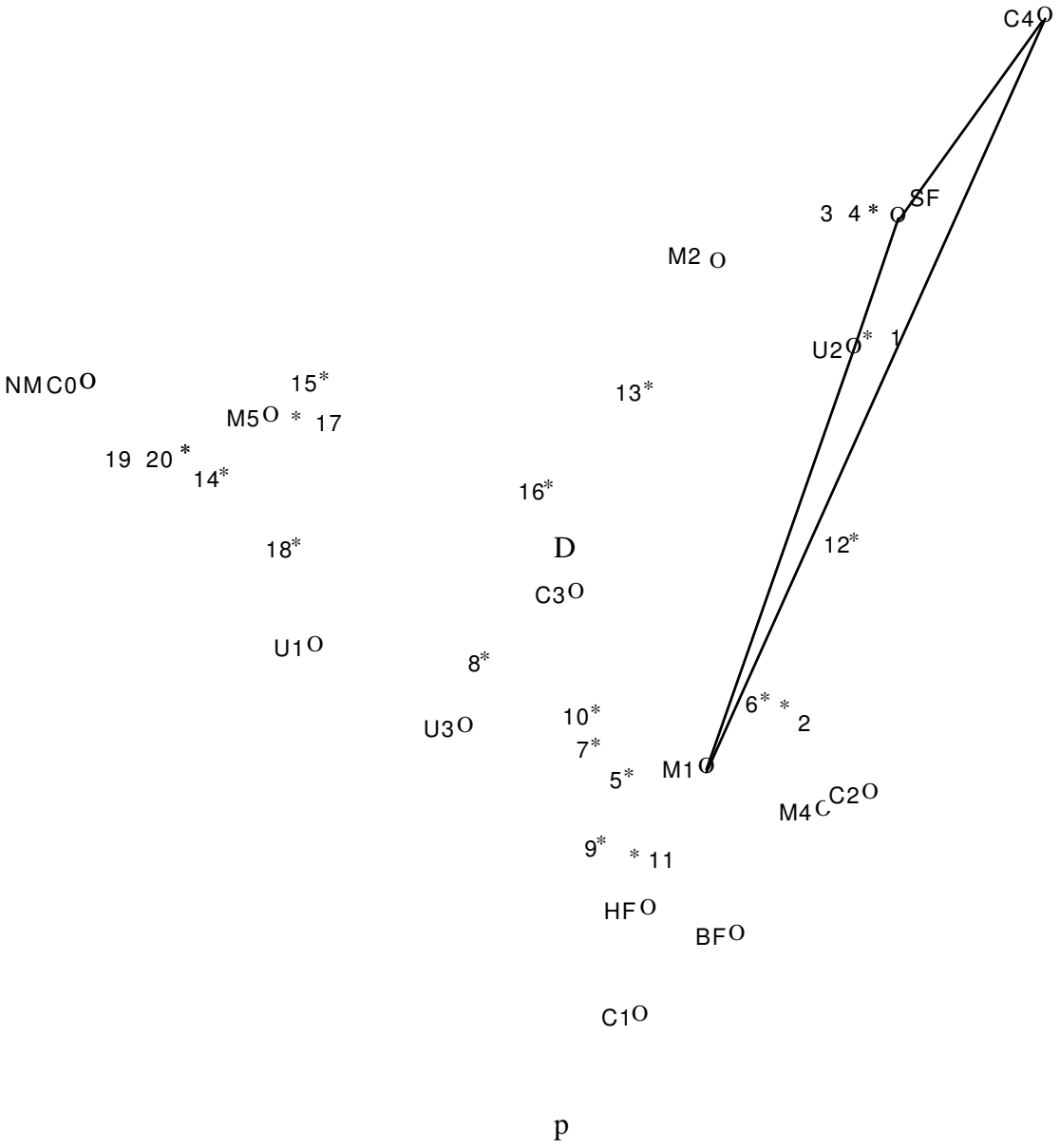

Figure 10: Multiple Correspondence Analysis. The biplot axes are now category-level-points represented by the letters; numbers refer to farms. This is an interpolative biplot and shows the interpolation of farm number 1.

Interpolation by the vector-sum method proceeds as before but rather than "extending by p times the centroid", the configuration of sample points may be multiplied by p and the centroid itself used for interpolation as is shown in Figure 10 which relates to environmental variables (Moisture, Management, Grass use and Manure use) recorded for 20 farms on the Dutch island of Terschelling. In this representation, every sample is nearest the CLPs for the category-levels it actually has than to any other CLPs. Thus, rather than the normal plane containing all points that predict x_k as in Figure 3, all points that predict a category-level lie an a region of space; where this region intersects the approximation space L defines

prediction-regions. The dominant concept is *nearness* which with axes is associated with projection but with CLPs induces neighbour regions.

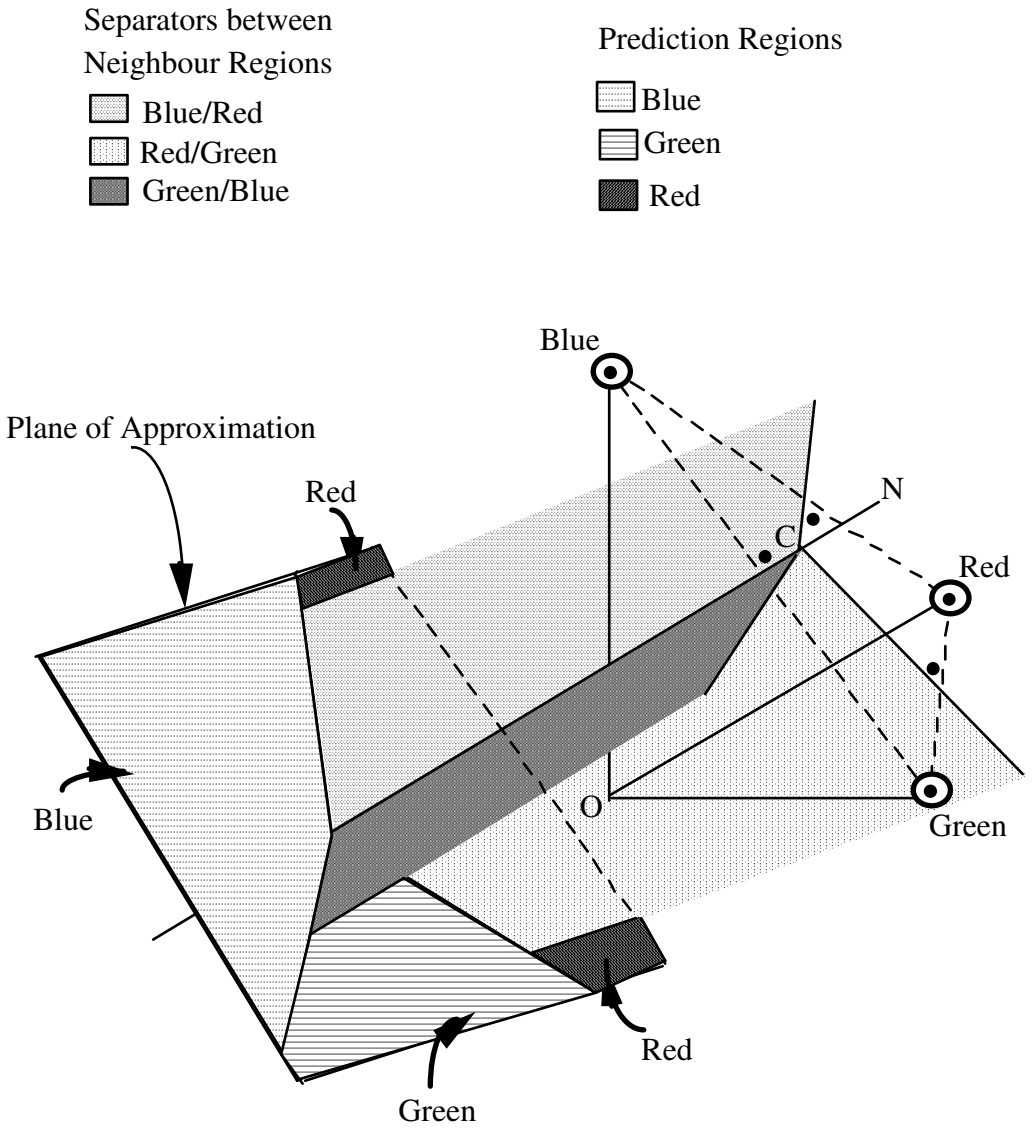

Figure 11: Diagram to explain CLPs, neighbour-regions and prediction regions.

Figure 11 illustrates the geometry for a categorical variables *colour* with three levels (*blue, green, red)* represented by the appropriately labelled CLPs. The planes bisecting the joins of all pairs of CLPs define regions of space within which one or other category-level is predicted - these are the *neighbour regions.* Where the neighbour-regions intersect the plane of approximation L, defines the *prediction regions*. In the figure, the prediction region for *red* is mostly hidden behind two of the bisecting planes.

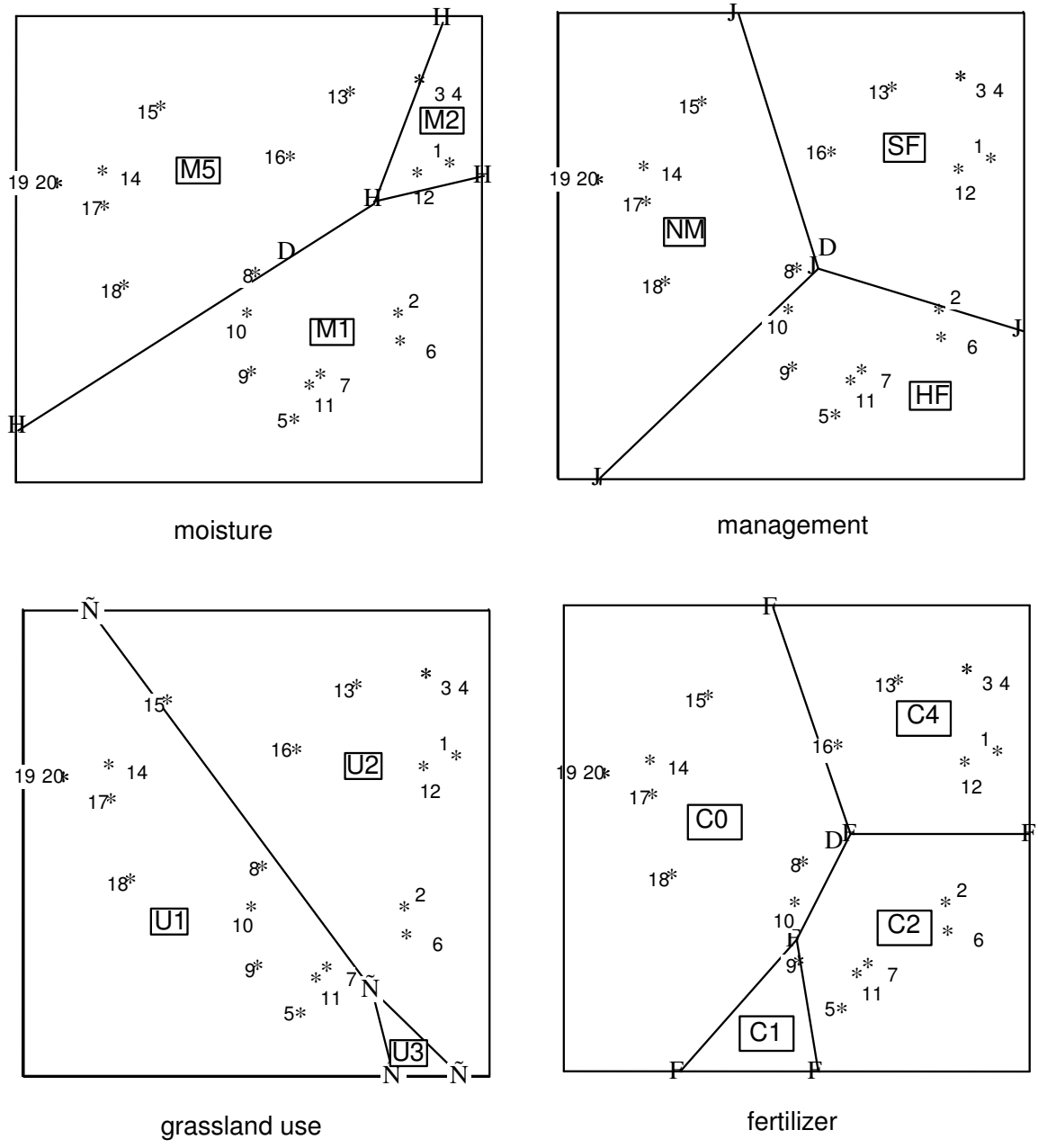

Figure 12: For categorical variables predictive biplot axes become prediction regions. The MCA prediction regions for four categorical variables are shown.

Some category-levels need not appear in the prediction-regions, being hidden behind the other regions. For example this has happened for the variable management which has four levels (SF, BF, HF, and NM) but BF does not occur in Figure 12.

Categorical variables need not necessarily be analysed by MCA. One could choose the CLPs differently, and perhaps better. Figure 13 shows the prediction regions when the CLPs are chosen as the rows of the unit matrix **I** where distance now corresponds to the Extended Matching Coefficient, which is a generalisation

of the Simple Matching Coefficient for binary variables. Figures 12 and 13 do not differ greatly but Figure 13 has the greater number of correct predictions. Whether this is a generally true or how to choose the CLPs and L to maximise the number of correct predictions are unresolved questions.

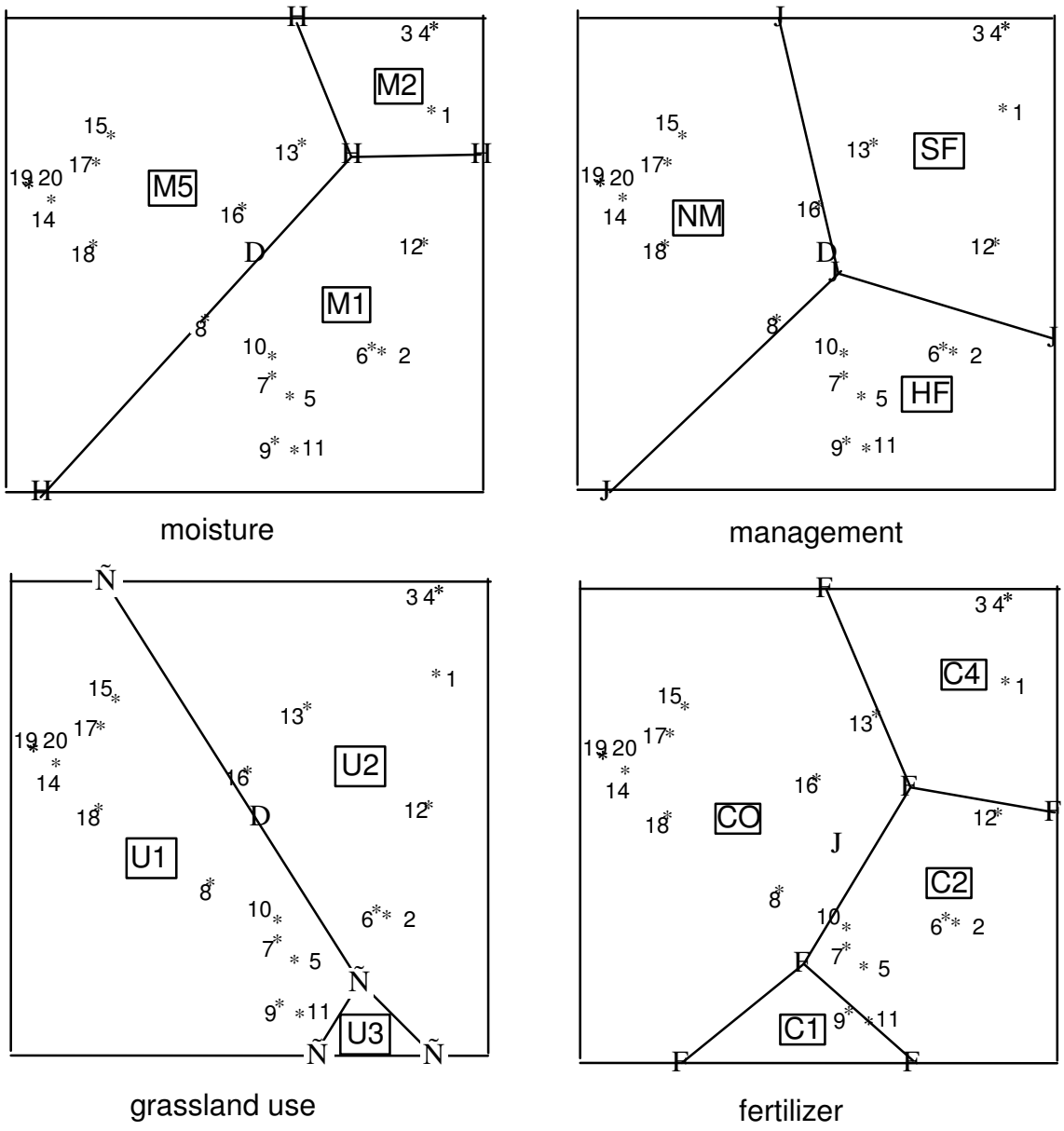

Figure 13: Prediction regions for the same four categorical variables as in Figure 12 but now using the Extended Matching Coefficient as the basis for calculating distance.

With quantitative variables, all the linear biplot axes may be shown conveniently on a single diagram. In Figures 12 and 13, each categorical variable is shown in a separate diagram; this is to avoid confusion. All four variables could be shown simultaneously and this is done in Figure 14 both for Chi-squared distance (MCA) and for the Extended Matching Coefficient. Such diagrams can be

useful for detecting associations between categorical variables but are best done in an interactive computing environment.

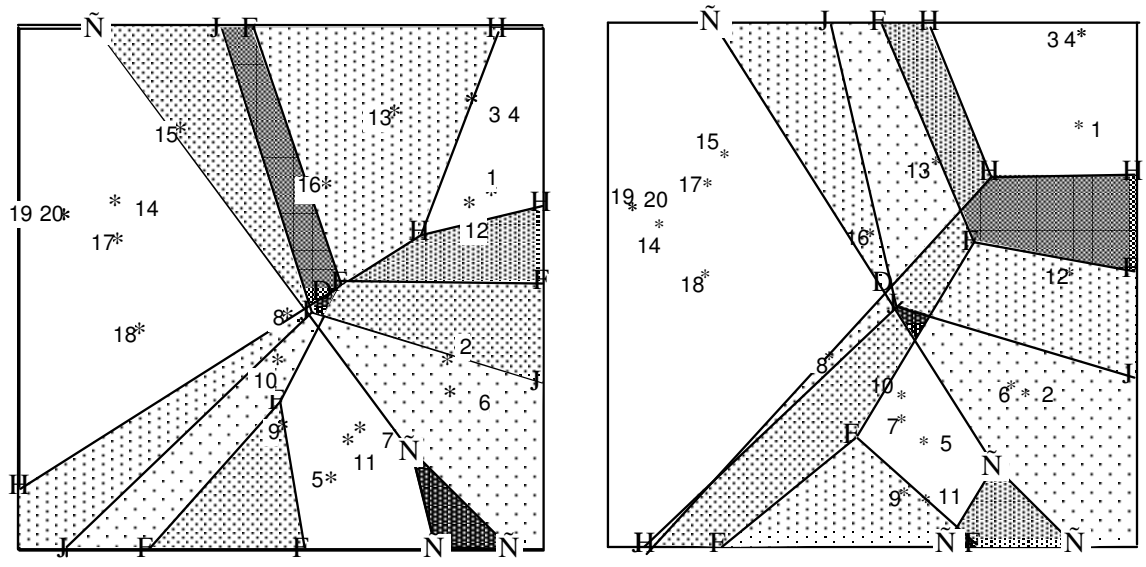

Figure 14: Superimposition of the prediction regions of Figures 12 and 13 respectively. These types of diagram are the equivalent of illustrating four quantitative variables by four linear biplot axes on the same diagram, e.g. Figure 8.

5 Canonical biplots

We are now concerned with data grouped between and within populations. Distance between populations is measured by Mahalanobis D^2. Both interpolative and predictive axes remain linear but, even in exact representations, have different directions as well as different scales. The reason from this follows from the following algebraic results. The data-matrix **X** now represents the group means and let **W** be the within-group dispersion matrix and $\mathbf{B} = \mathbf{X}'\mathbf{X}$ be the within group dispersion. Then we require solutions to the two-sided eigenvalue problem:

$\mathbf{BL} = \mathbf{WL}\Lambda$ normalised so that $\mathbf{L}'\mathbf{WL} = \mathbf{I}$.

The *r*-dimensional weighted least-squares approximations to **X** is given by $\mathbf{X}_r = \mathbf{X}\mathbf{L}_r\mathbf{L}^r$ where $\mathbf{L}_r$ represents the first *r* columns of **L** and $\mathbf{L}^r$ represents the first *r* rows of $\mathbf{L}^{-1}$. This a simple way of representing the generalised Eckart-Young theorem for weighted least-squares approximation to a matrix. The rows of $\mathbf{L}_r$ give the directions of the interpolative axes and the columns of $\mathbf{L}^r$ give the directions for prediction. In PCA $\mathbf{W} = \mathbf{I}$ and **L** is orthogonal so that then $\mathbf{L}^{-1} = \mathbf{L}'$ and the interpolative and predictive axes are the same; in the general case the two matrices differ. Figure 15 shows interpolation which remains by calculating vector-sums while Figure 16 shows prediction, which remains by orthogonal projection. Similar biplots exist for canonical correlation.

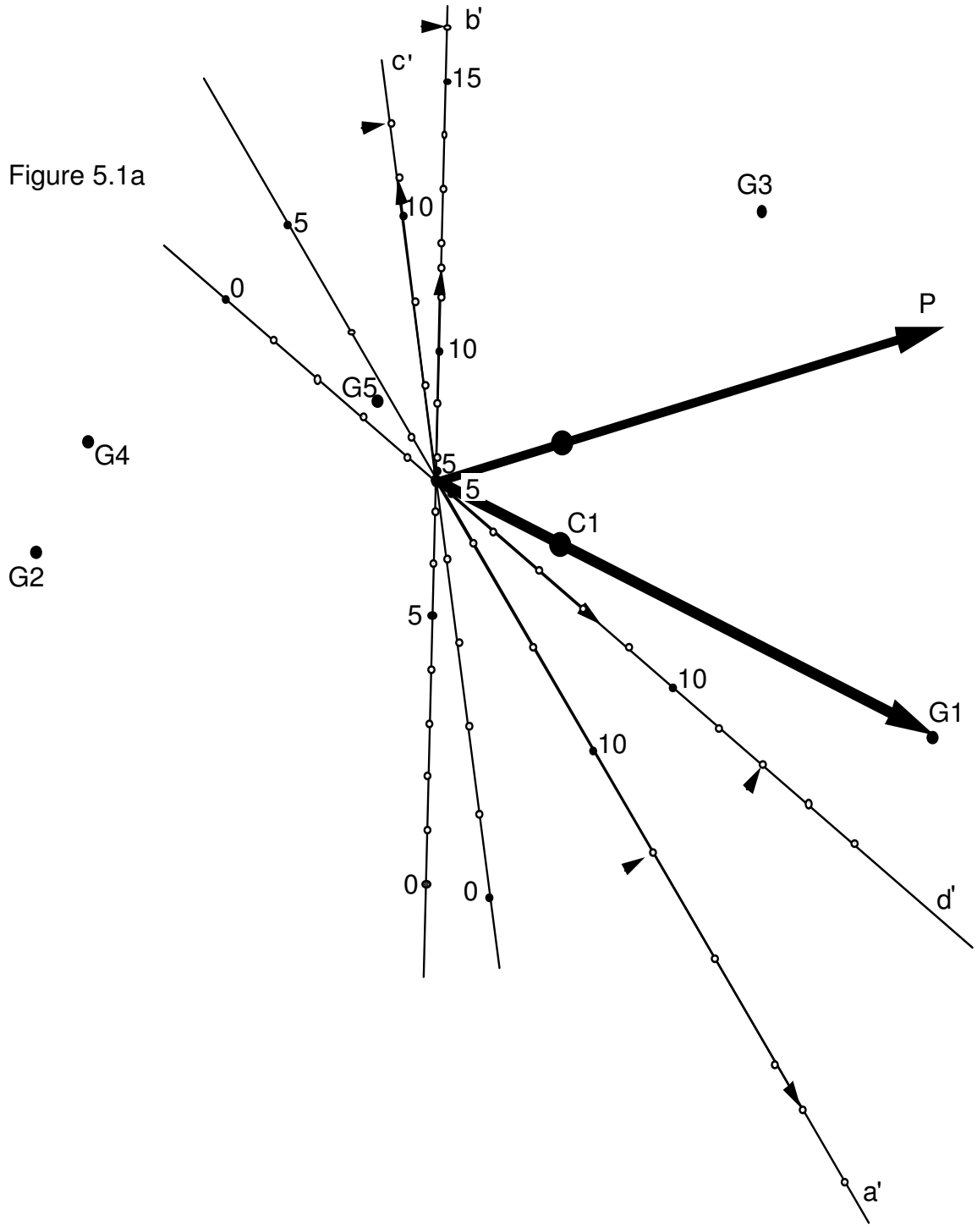

Figure 15: Interpolative biplot axes for Canonical Variate Analysis.

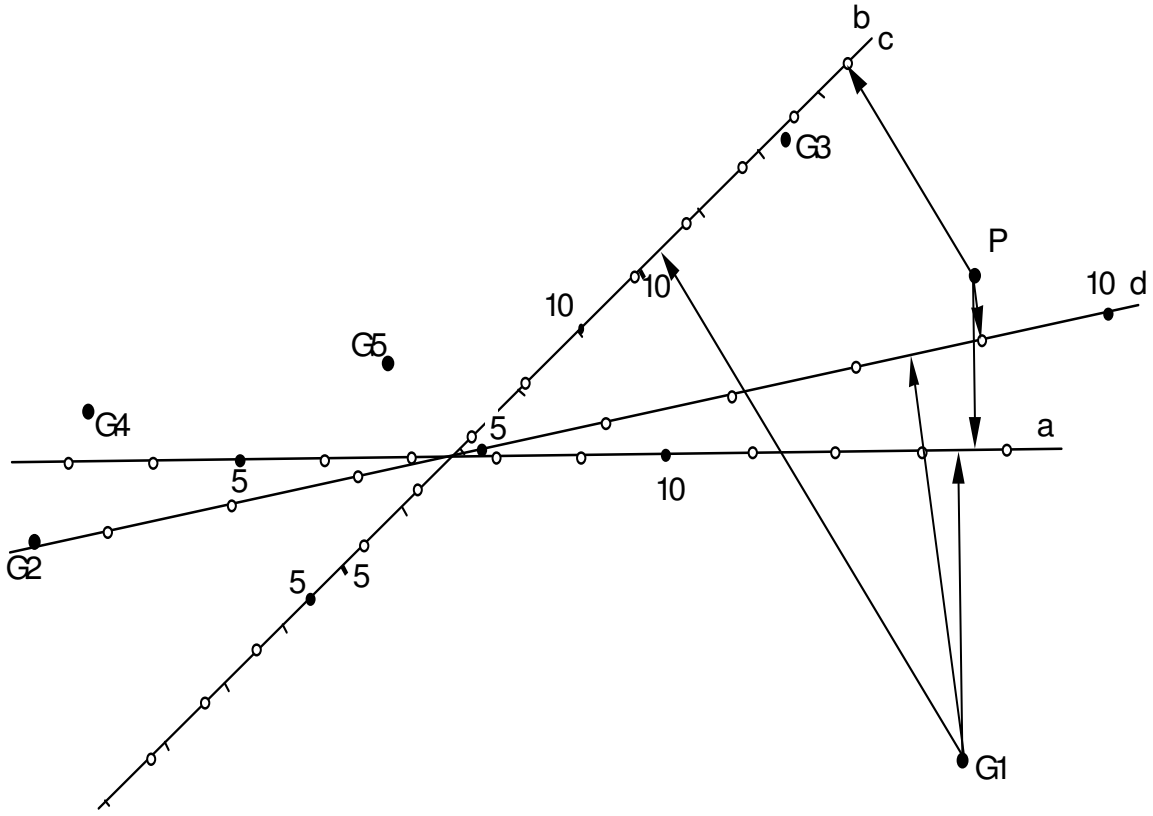

Figure 16: Predictive biplot axes for Canonical Variate Analysis.

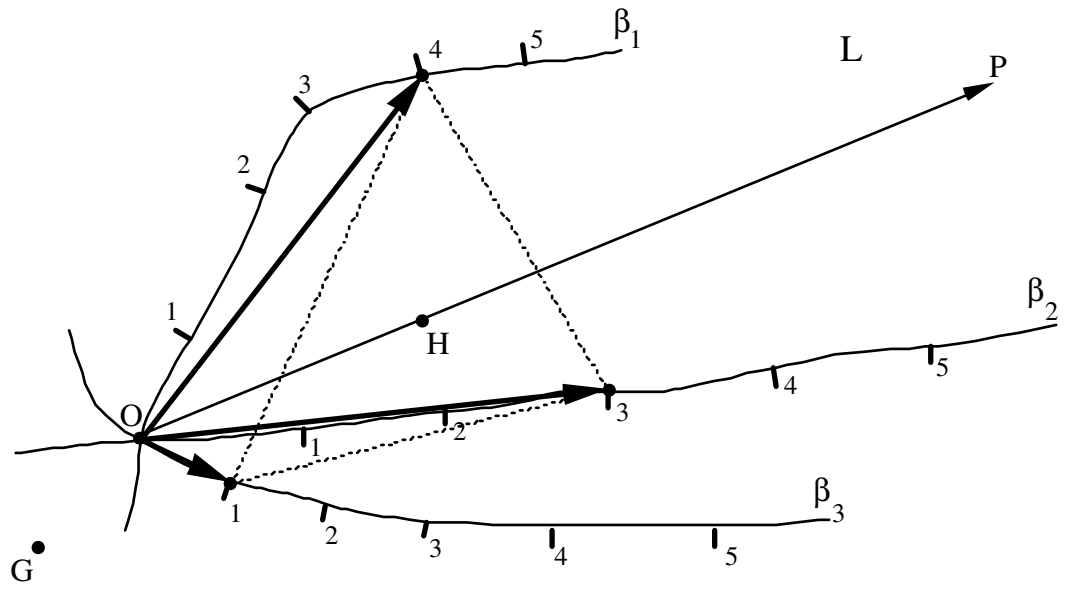

Figure 17: Interpolative non-linear biplot axes, With classical scaling/principal coordinates analysis; the vector-sum method remains valid.

6 Non-linear and generalised biplots.

The biplot ideas may be extended to other forms of distance. Starting with the data-matrix **X** we define a set of Euclidean embeddable distances d_{ij} which are generated by a set of coordinates **Y**. The original axes may be represented in the same space but the axes become non-linear trajectories and markers for equal steps in the original variables become unequally spaced. This gives an exact representation in which coordinates remain given by the *nearest* markers on the trajectories. Approximations in L may be found by PCO and then interpolation by the vector-sum method remains valid as illustrated in Figure 17, where $d_{ij}^2 = \sum_{k=1}^{12} \log(x_{ik} - x_{jk})$. Even when Euclidean embeddability is impossible, the method remains useful. Predictive non-linear biplots are also available. The geometrical basis for prediction is similar to that of Figure 4 but the axis ξ_k is now non-linear. This induces non-linear biplot axes β_k in L. Prediction is either by *circular projection*, which is numerically simple to compute but which is somewhat cumbersome to use, or by normal projection, which requires the numerical solution of differential equations but which is easy to use.

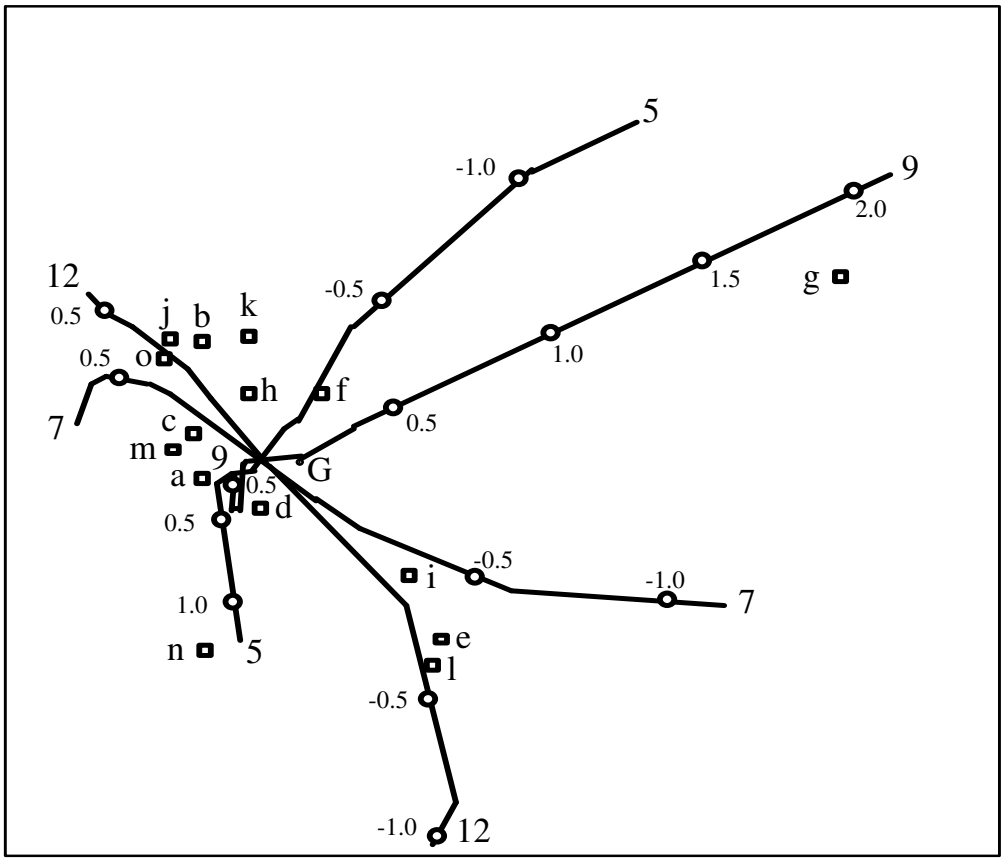

Figure 18: An example of an interpolative non-linear biplot. Four variables of amounts of trace elements at 15 sites in Glamorganshire.

Generalised biplots extend non-linear biplots to include categorical as well as quantitative variables. The concept of nearness remains the basis of interpretation. Numerical variables generate non-linear trajectories while categorical variables generate CLPs. With PCO, explicit algebraic results are available for the CLPs and their projections. Figure 18 shows vector-sum interpolation with a mixture of categorical and quantitative variables. For prediction, the concept of prediction-regions remains in force.

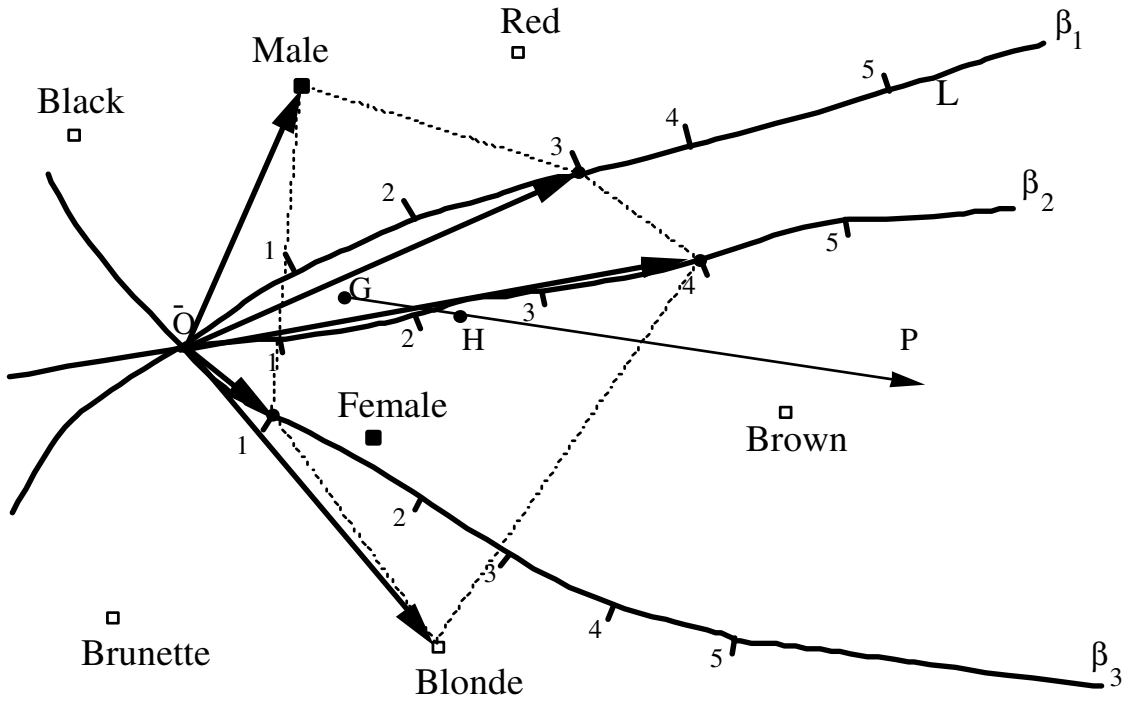

Figure 19: A generalised interpolative biplot showing the vector-sum method in use with categorical variables represented by CLPs simultaneously with quantitative variables represented by non-linear axes.

7 Further developments

The above has given the briefest of overviews. We have not covered biplots for two-way tables that include biadditive models for quantitative variable and correspondence analysis of contingency tables. In the two-way table context, biplots may be found after adjusting for one or more covariates; this has close links with redundancy analysis. Neither have we discussed biplots for special classes of matrix (e.g. covariance and correlation matrices) or biplots for symmetric and skew symmetric matrices. Associating biplots with non-metric scaling is straightforward and in some cases leads to linear axes with irregularly

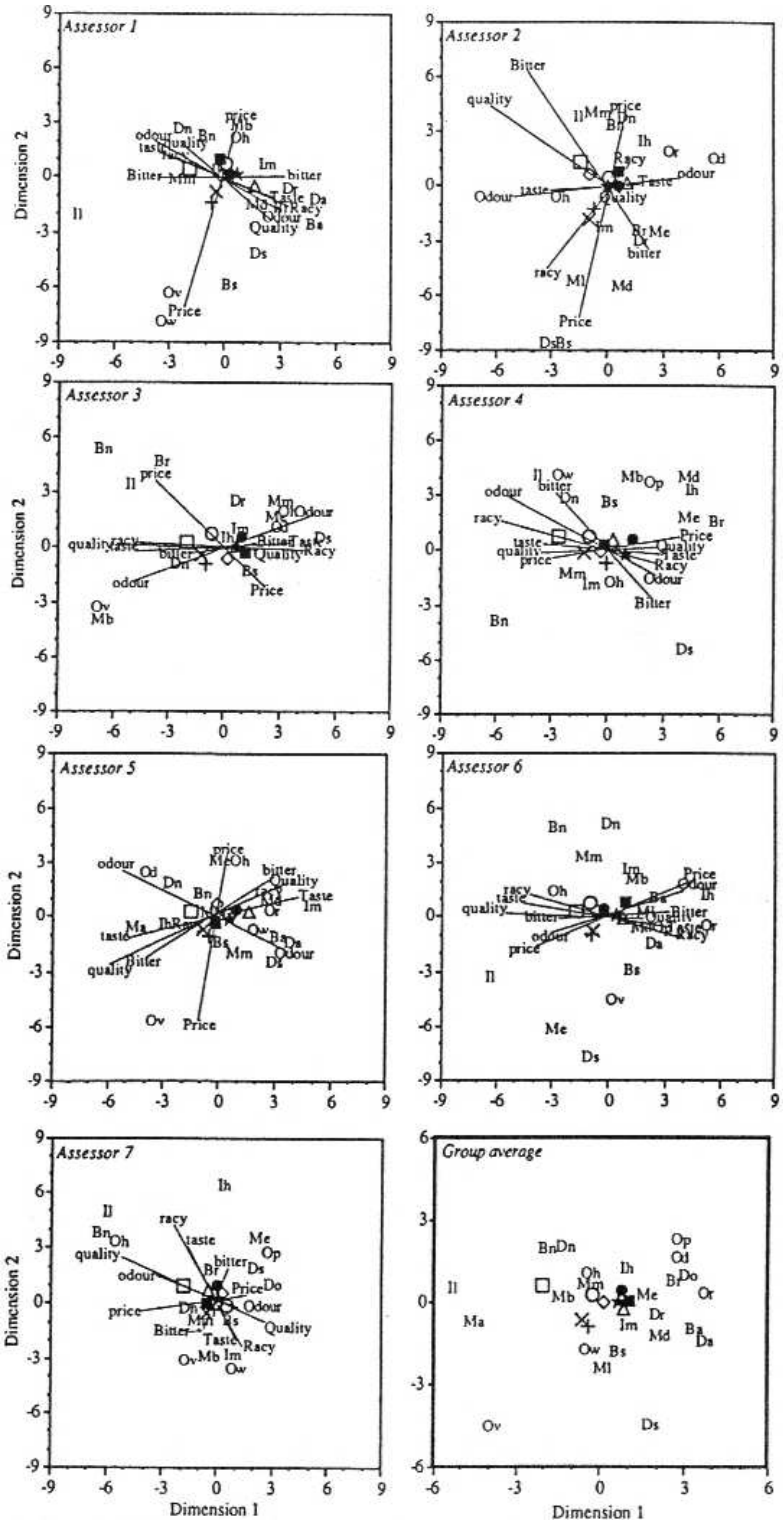

Figure 20: Two dimensional representations of the configurations for seven assessors rotated to best fit their group average. The rotations were done in the maximal space of 26 dimensions. Everything is referred to the principal axes of the group average. Quantitative variables are labelled at their higher values. Nine brands of coffee are denoted by the symbols $\Delta * + \times \square \circ \Diamond \blacksquare \bullet$.

spaced markers that are necessarily monotonic scales only for ordered categorical variables (see Gower, Meulman and Arnold, 1999). When all the variables are of the ordered categorical type, we may represent the prediction regions in terms of what look like ordinary linear biplot axes; then the simultaneous representation of all the categorical variables presents no problem (Gower, J.C. and Ngouenet, R.F., 1998). Further information and applications of prediction regions are given by Gower and Harding (1998). The concept of CLPs with associated neighbour and prediction regions has been developed further to suggest new methods of MDS that optimise correct prediction rates but these seem to require difficult combinatorial algorithms (Gower, 2002). The ideas may also be extended to three-mode analyses - at least in some cases. Figure 20 illustrates the possibilities, showing generalised interpolative biplots for a Generalised Procrustes Analysis of seven assessors of packaging for nine brands of coffee; there are 11 variables of which six are quantitative contributing to Pythagorean distance and the remaining five are categorical, contributing to the extended matching coefficient. The lower right-hand diagram refers to the group-average, obtained by averaging the other seven diagrams, the averages of the linear axes for quantitative variables has been omitted.

As with most modern statistical developments confident interpretation depends on experience and experience depends on the availability of user-friendly software. Most biplots given here have been produced by Genstat 5. Many are computationally straightforward but good graphics properly labelled and with sensible scales are hard to produce.

References

[1] Gower, J.C. and Hand, D.J. (1996): *Biplots*. Monographs on Statistics and Applied Probability, **54**. London: Chapman and Hall, 277 p.

[2] Gower, J.C. and Ngouenet, R.F. (1998): Some new types of biplot. *Proceedings of the Fourth Sensometrics Conference*, Copenhagen. 60-63.

[3] Gower, J.C. and Harding S.A. (1998): Prediction regions for categorical variables. In J. Blasius and M.J. Greenacre (Eds): *Vizualisation of Categorical Variables*. London: Academic Press, 405-419.

[4] Gower, J.C., Meulman, J.J., and Arnold, G.M. (1999): Non-metric linear biplots. *Journal of Classification*, **16**, 181-196.

[5] Gower, J.C. (2002). Categories and Quantities In S. Nishisato, Y. Baba, H. Bozdogan and K. Kanefuji (Eds.): *Measurement and Multivariate Analysis*. Tokyo: Springer, 1-12.