BIPLOT GEOMETRY

John C. Gower

Department of Data Theory

University of Leiden

BIPLOT GEOMETRY

J. C. Gower
Department of Data Theory
University of Leiden

Summary

Biplot axes are here interpreted as coordinate axes. For linear biplots, interpretation must be in terms of non-orthogonal axes and it turns out that there is a fundamental distinction in how these axes are scaled, depending on whether they are to be used for interpolation or prediction. Interpolation is expressed in terms of vector-sums; prediction in terms of what is here named back-projection, in which the concept of nearness is fundamental. The methodology extends to non-linear biplots; in these, linear axes are replaced by non-linear trajectories but now, rather than just different scales on the same biplot axes, separate sets of trajectories are needed for the interpolation and prediction operations. The methodology also embraces categorical variables; in these, axes are replaced by a simplex of category-level points. Interpolation remains by vector-sums and prediction by back-projection but nearness now plays a more overt part and leads to the consideration of neighbour-regions.

1 Introduction

One of the most basic and frequently used statistical methods is to plot a scatter diagram showing the pattern of relationships between a set of samples, on which there are two measured variables x and y, say. One may go on to fit a curve to this scatter, or one may be interested in the possible clustering of samples, or one may be interested in outliers, or one may be interested in collinearities or other regularities. The fundamental tool here is representation with respect to orthogonal Cartesian coordinate axes with the scales of measurement (perhaps in some normalised form) marked along the two axes.

The most simple multivariate extension is to represent samples on which there are p measured variables, relative to p orthogonal axes. Because p-dimensional Euclidean space cannot easily be visualised, it is natural to introduce the additional step of approximating the p dimensional relationships in few, usually two, dimensions. There are many ways of approximating but let us consider, at least initially, one of the most popular - principal components analysis. Here the p-dimensional scatter of the samples is approximated by their scatter in an r-dimensional sub-space, obtained by orthogonal projection of $\mathfrak{R}_p$ onto $\mathfrak{R}_r$, chosen to minimise the sum-of-squares of the residuals orthogonal to $\mathfrak{R}_r$. The approximation of the variables is given by biplot axes (Gabriel, 1971) which are the p vectors through the origin, G, obtained as the orthogonal projections of the Cartesian axes onto $\mathfrak{R}_r$. Of course, p orthogonal axes cannot be represented orthogonally in fewer than p dimensions, so that in an r-dimensional approximation most, if not all, of the biplot axes must be oblique. Furthermore, to represent r-dimensional coordinates relative to p (oblique) axes, where $r < p$, introduces an element of superfluity which needs consideration. Gower (1991) has stressed the usefulness of regarding the biplot axes as coordinate axes, marked with the scales of measurement of the original variables in the same way as were the original orthogonal axes. Although this is not the usual basis for interpretation used with biplots, nevertheless it is the one that is developed in the following. This approach is taken in the belief that interpretation in terms of coordinate axes offers a key to the understanding of several forms of biplot. There is more to their proper understanding than might at first be thought. I believe that full understanding can come only from a thorough exploration of the underlying geometry and this is the theme of section 2. Indeed, this preliminary geometrical study is essential when one considers the extension of classical linear biplots to the non-linear form (Gower and Harding, 1988), which admits the use of general distances among quantitative variables, and to generalised biplots (Gower, 1991), which further admits categorical variables. The geometry of these extensions is discussed in sections 3 and 4. In their turn, these generalisations throw more light on the special case of the linear biplot.

1.1 Notation

Matrices are printed in bold capitals, vectors in bold lower-case and scalars in italics. The sizes of matrices are given initially but thereafter sizes are by implication. By convention, vectors are column-vectors when they refer to directions and row-vectors when they refer to coordinates - the distinction is artificial because the end-point of a direction may be regarded as a set of coordinates and the coordinates of a point have a bearing from the origin; nevertheless the distinction is found convenient. A vector of units is written $\mathbf{e}$ and $\mathbf{e}_k$ is a unit vector on the kth axis. Vector-spaces are written in curly capitals $\mathfrak{R}$ or $\mathfrak{R}_p$, where the suffix, when present, gives dimensionality. Coordinate axes, not necessarily

linear, are named by lower-case greek letters (e.g. ξ_k) with one suffix, and a coordinate position on such an axis has a label X_i and a marker x_{ik}, giving the value of the variable at the point labelled X_i on the kth axis; the suffix k will be dropped when there is no ambiguity. Notionally, markers exist at all points of ξ_k but, in practice, physical markers will be present at only few points, usually spaced at equal intervals of the values of the variable. A consistent notation ($\mathbb{X}$, X_i, x_{ik}, ξ_k) linking all these concepts would have been desirable but was found to be impracticable, partly because sub-spaces would then have two, or more, nomenclatures but also because in the important one-dimensional case, $\mathbb{X}$ and ξ_k become synonyms.

The origin of $\mathfrak{R}$ has already been labelled G. This unusual notation is used because, in the following, the origin is at the centroid of a set of points. The symbol O is reserved for a point representing the sample means of a set of variables; O and G coincide in components analysis, but not otherwise. Centroids occurring in other contexts are labelled H.

1.2 Algebra of Linear Biplots

The methodology of linear biplots may be expressed very simply by considering the $n \times p$ data-matrix $\mathbf{X}$, assumed centred at its mean, so that $\mathbf{e}'\mathbf{X} = \mathbf{0}$, and using the singular value decomposition:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}' \qquad (1)$$

then plotting the rows of $\mathbf{U}\mathbf{\Sigma}$ to give n p-dimensional coordinates for the samples and the p rows of $\mathbf{V}$, each of which gives the p-dimensional coordinates of one point on a corresponding biplot axis. It is assumed that the singular values on the diagonal of $\mathbf{\Sigma}$ are presented in non-increasing order. In r-dimensional approximations, $\mathbf{V}$ and $\mathbf{\Sigma}$ are replaced by their first r columns, $\mathbf{V}_r$ and $\mathbf{\Sigma}_r$. The approximation of $\mathbf{X}$ in $\mathfrak{R}_r$ is given by its orthogonal projection $\mathbf{X}\mathbf{V}_r = \mathbf{U}\mathbf{\Sigma}\mathbf{V}'\mathbf{V}_r = \mathbf{U}\mathbf{\Sigma}_r$; these coordinates are often termed the r-dimensional principal component scores. Any sample $\mathbf{x}$, not necessarily one whose coordinates are given by a row of $\mathbf{X}$, can be projected into $\mathfrak{R}_r$ by evaluating $\mathbf{x}\mathbf{V}_r$. In particular, a value of one unit of measurement on the kth Cartesian axis becomes $\mathbf{e}_k\mathbf{V}_r$ on the kth biplot axis, and should be so marked to give the correct scale. Thus if $\mathbf{x} = \sum_{k=1}^{p} x_k\mathbf{e}_k$, its projected position in $\mathfrak{R}_r$ is given by $(\sum_{k=1}^{p} x_k\mathbf{e}_k)\mathbf{V}_r = \sum_{k=1}^{p} x_k(\mathbf{e}_k\mathbf{V}_r)$. This means that to interpolate $\mathbf{x}$, one need only take the vector-sum of the markers $(x_1,x_2,...,x_p)$ on the biplot axes.

When P_i corresponds to one of the n samples, prediction of the values $(x_{i1},x_{i2},...,x_{ip})$ to be associated with a given point P_i in $\mathfrak{R}_r$ is also a common requirement. Prediction is given by the Eckart-Young (1936) theorem (see section 2.3), which gives the least-squares rank r inner-product approximation to (1):

$$\hat{\mathbf{X}} = \mathbf{U}\mathbf{\Sigma}_r\mathbf{V}_r',$$

so that $\hat{\mathbf{x}}_{ik} = (\mathbf{u}_i\mathbf{\Sigma}_r)\mathbf{v}_k'$. That is, we have to project the ith sample-point of $\mathfrak{R}_r$ onto the kth biplot axis, β_k, and multiply by the length of $\mathbf{v}_k$, whose elements are the first r columns of the kth row of $\mathbf{V}$. The multiplication can be avoided by marking the length of $\mathbf{v}_k$ as a unit point on the kth axis, giving a scale for prediction that differs from that described above

for interpolation. Relative to this new scale, $\hat{x}_{ik}$ may be read off immediately against the prediction-marker nearest the projection of P_i onto β_k The same method applies even when P_i does not correspond to an observed sample. Figure 1 illustrates the situation.

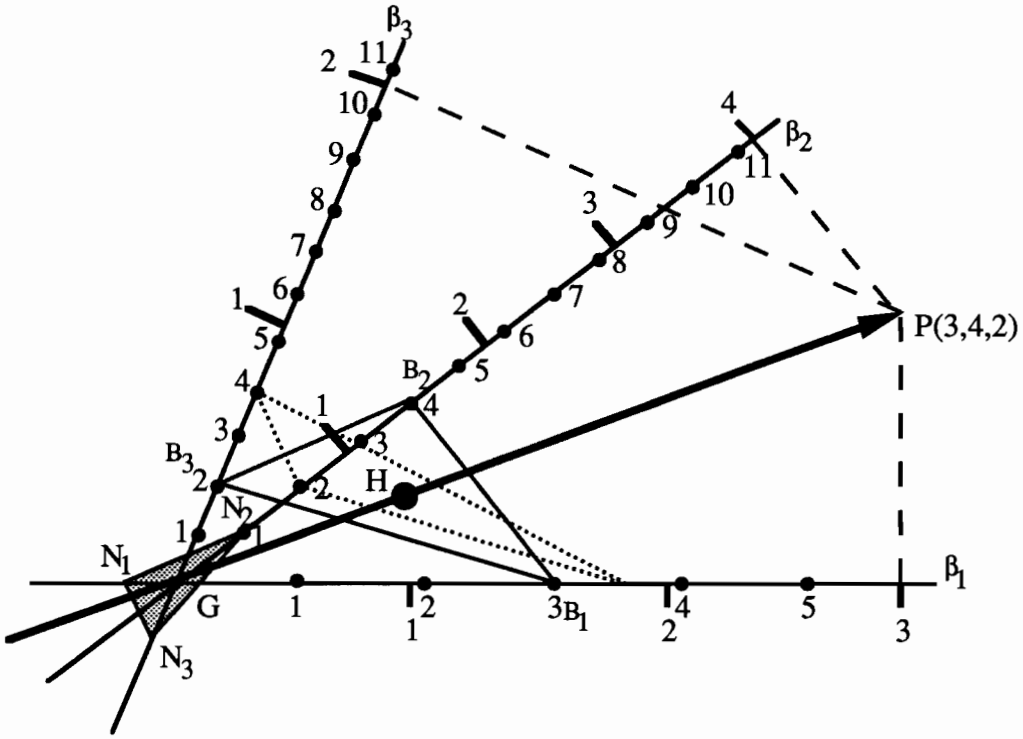

Figure 1. The biplot axes β_1,β_2 and β_3 have circular markers for the prediction scales and vertical markers for the interpolation scales. The figure shows (i) the interpolation of a point (3,4,2) at P through the centroid H of B_1, B_2, B_3, (ii) the predictions associated with P, (iii) the null-point with coordinates labelled N_1, N_2, N_3 and (iv) the interpolation of the point $(3\frac{1}{2}, 2, 4)$, also with centroid H and hence also at P.

Figure 1 prompts the following remarks, some of which may not be fully comprehended until section 2 has been read:

(1) The unit of scale for interpolation is always smaller than the unit of scale for prediction, because projection reduces length.

(2) Scales for different variables differ, even if they were equal before projection onto $\mathfrak{R}_r$.

(3) Every point P_i of $\mathfrak{R}_r$ generates a unique set of predictions.

(4) In the figure, the values (3,4,2), denoted by the triangle B_1,B_2,B_3, is interpolated at P by the vector-sum rule; P is also predicted as (3,4,2), so the two scales are consistent for these values. This is because P must actually lie in $\mathfrak{R}_r$. The values $(3\frac{1}{2},2,4)$, marked by the triangle with dashed lines, also interpolate into P, so the scales are not consistent for these values. This is because these are the coordinates of a point that is outside $\mathfrak{R}_r$. The small shaded triangle N_1,N_2,N_3 is consistent for both interpolation and prediction at the origin; any multiple of its coordinates may be added to any consistent set of coordinates to give an interpolation that is not consistent with prediction. This is a manifestation of many points of $\mathfrak{R}_p$ projecting into a single point of $\mathfrak{R}_r$. In the example shown, the coordinates of the dotted triangle are those of B_1,B_2,B_3 less twice the coordinates of N_1,N_2,N_3. With respect to the interpolation scales, the more discrepant the values of an inconsistent set of coordinates, as compared to those of its consistent counterpart, the further is the sample-point from $\mathfrak{R}_r$. Because in a good approximation all points should be close to $\mathfrak{R}_r$, residuals should be small and most values occurring in a sample should be close to being consistent for interpolation and prediction.

(5) The maximum and minimum values of $\mathbf{x}_k$ in the sample determine natural limits for the scales as plotted. Just as with ordinary x-y plots, one would choose the lengths of the axes to accommodate comfortably the values that had actually occurred or might reasonably be expected to occur. Axes with short lengths, relative to other axes, play little part in the approximation and might be deleted.

(6) A unit of one standard deviation is often marked on biplot axes. This raises no new problems and can be useful when $\mathbf{X}$ is believed to be a sample from some well-behaved distribution. Often, it might be better to mark on each biplot axis β_k, the actual values occurring for the n samples in the kth variable, thus giving a flat histogram indicating skewness, outliers, or the degree of non-uniform sampling.

In this section, some of the properties of biplots have been rehearsed. The distinction between interpolation and prediction has been shown to induce two scales; the interaction between the degree of approximation and the two scales has also been demonstrated, and it has been shown that useful high-dimensional information may be preserved in two dimensions. These types of consideration underly much of the following discussion.

2 The Geometry of Linear Biplots

It is convenient to begin by reexamining the fundamental representation given by orthogonal Cartesian axes and to consider the modifications required for oblique axes. Because much of the required geometry does not depend on the optimal approximation described in section 1, in this section the r-dimensional sub-space $\mathfrak{R}_r$ is arbitrary, unless otherwise specified.

2.1 Coordinate Axes

Cartesian coordinate axes are so familiar, that some of the implications of their use and of the different ways in which they may be characterised, can be overlooked. Consider coordinates $P(x_1,x_2)$ relative to two orthogonal Cartesian axes. Conventionally, this determines a point P in two-dimensional space by moving, from an origin, x_1 units along a first axis in an easterly direction and then x_2 units parallel to a second axis in a northerly direction. Alternatively we may regard the point arrived at as the vector-sum $x_1\mathbf{e}_1 + x_2\mathbf{e}_2$, as described in section 1 for interpolating a point into a biplot representation. A third characterisation, the normal-plane method, regards P as being at the intersection of the normal at the marker x_1 on the first axis with the normal at x_2 on the second axis, relating to prediction as described in section 1. These three characterisations easily generalise to multiple orthogonal axes.

With two oblique axes, vector-sums become the parallel axis method, where $P(x_1,x_2)$ is obtained by completing the parallelogram whose other three vertices are $(x_1,0)$, $(0,x_2)$ and the origin. Erecting normals on oblique axes determines a different position for (x_1,x_2). Thus with non-orthogonal axes, these two characterisations diverge and, anticipating future developments, this suggests that the underlying geometries of interpolation and prediction differ, as indeed we have already seen in section 1 to the extent that the measurement scales associated with these two objectives differ, even with classical linear biplots.

2.2 A General Result

The result of this section will be presented in a fairly abstract way. Its relevance to understanding biplots will become clear in following sections. Suppose $\mathcal{L}$ and $\mathcal{M}$ are two linear sub-spaces of $\mathcal{R}_p$. Initially it is assumed that both contain the origin. Suppose $\mathbf{x}$ is a given point of $\mathcal{M}$, then all points in the space $\mathcal{N}$ that is normal to $\mathcal{M}$ at $\mathbf{x}$ will be said to predict the value $\mathbf{x}$ (the motivation for this definition is the normal plane method for defining the coordinates of a point, in which all points in the space normal to a coordinate axis at a marker x have coordinate x). When $\mathcal{N}$ intersects with $\mathcal{L}$ then $\mathcal{N} \cap \mathcal{L}$ contains all points in $\mathcal{L}$ that predict $\mathbf{x}$; among these points there will be $\mathbf{y}$, which is nearest $\mathbf{x}$. This section is concerned with the algebraic expression for the transformation of $\mathbf{x}$ to $\mathbf{y}$ and in its geometrical properties. The geometry is exhibited in Figure 2.

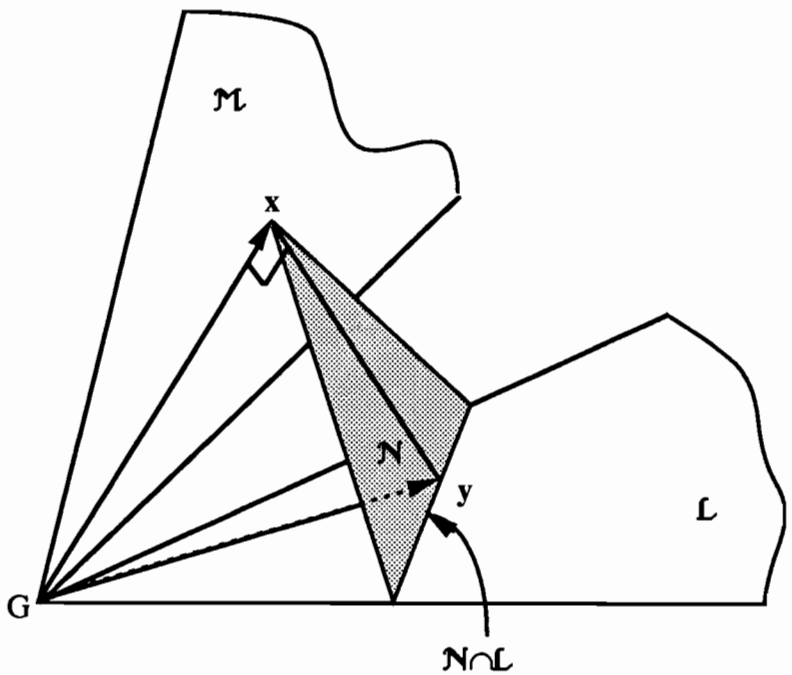

Figure 2. $\mathcal{N}$ is normal to $\mathcal{M}$ at $\mathbf{x}$. $\mathbf{y}$ is the point in $\mathcal{N} \cap \mathcal{L}$ that is nearest $\mathbf{x}$. $\mathbf{y}$ is termed the back-projection of $\mathbf{x}$ in $\mathcal{L}$.

Suppose that $\mathcal{L}$ is spanned by a set of r independent column-vectors given in the $p \times r$ matrix $\mathbf{L}$ and $\mathcal{M}$ by s independent column-vectors given in the $p \times s$ matrix $\mathbf{M}$. Without loss of generality, and with the benefit of improved algebraic simplicity, $\mathbf{L}$ and $\mathbf{M}$ are taken to be orthonormal throughout. That $\mathbf{y}$ lies in $\mathcal{L}$ and $\mathbf{x}$ lies in $\mathcal{M}$ is expressed by:

$$\mathbf{yKK'} = \mathbf{0} \text{ and } \mathbf{xNN'} = \mathbf{0}. \qquad (2)$$

where $\mathbf{KK'} = \mathbf{I} - \mathbf{LL'}$ and $\mathbf{NN'} = \mathbf{I} - \mathbf{MM'}$, where $\mathbf{K}$ and $\mathbf{N}$ are assumed to give orthonormal bases for the complementary spaces of $\mathcal{L}$ and $\mathcal{M}$, respectively. In particular, the columns of $\mathbf{N}$ may be chosen to be an orthonormal basis for $\mathcal{N}$. Because $\mathbf{y}$ is nearest $\mathbf{x}$, then $\mathbf{x}$ is an orthogonal projection of $\mathbf{y}$ in $\mathcal{M}$, i.e.:

$$\mathbf{x} = \mathbf{yMM'} \qquad (3)$$

which is consistent with the second equation of (2). Assuming that $s \leq r$, a solution (see appendix A) for $\mathbf{y}$ in terms of $\mathbf{x}$, which accommodates the general case when $\mathcal{L}$ and $\mathcal{M}$ are disjoint spaces, is:

$$\mathbf{y} = \mathbf{x}(\mathbf{I} - \mathbf{K}(\mathbf{K'NN'K})^{-1}\mathbf{K'NN'}). \qquad (4)$$

Clearly (4) satisfies (2) and (3). We term (4) the *back-projection* of $\mathbf{x}$ in $\mathcal{L}$. After some algebraic manipulation and recalling that $\mathbf{KK'} = \mathbf{I} - \mathbf{LL'}$, (4) may be written in terms of the matrices $\mathbf{L}$ and $\mathbf{M}$ as:

$$\mathbf{y} = \mathbf{x}(\mathbf{I} + \mathbf{KK'M}(\mathbf{M'LL'M})^{-1}\mathbf{M'})\mathbf{LL'}. \qquad (5)$$

When, as in Figure 2, $\mathcal{L}$ and $\mathcal{M}$ are not disjoint, (4) and (5) simplify to:

$$\mathbf{y} = \mathbf{xM}(\mathbf{M'LL'M})^{-1}\mathbf{M'LL'}. \qquad (6)$$

The results (4), (5) and (6) are central to the following development.

The assumption that $\mathcal{L}$ and $\mathcal{M}$ both contain the origin may be relaxed. Suppose the origin G is in $\mathcal{L}$ but that $\mathcal{M}$ is offset by a vector $\mathbf{q}$, defined to be the projection of G onto $\mathcal{M}$. Then (2) and (3) become:

$$\mathbf{yKK'} = \mathbf{0},\ \mathbf{xNN'} = \mathbf{q} \text{ and } \mathbf{xMM'} = \mathbf{yMM'},$$

which, it may be easily verified, remain satisfied by (4), and hence by (5). Appendix A shows that (6) remains satisfied when $\mathbf{qL} = \mathbf{0}$, which includes the important case of a common origin ($\mathbf{q} = \mathbf{0}$). For most of this paper this condition is satisfied but in sections 3.3 and 4.3, $\mathbf{qL} \neq \mathbf{0}$ and the simplification (6) is then unavailable as a back-projection but remains useful with a different interpretation.

2.3 The Eckart-Young Theorem

Consider the special case where $\mathcal{M}$ is the kth Cartesian coordinate axis, so now $\mathcal{M}$ is a synonym for ξ_k. and is one-dimensional. $\mathcal{L}$ remains any linear sub-space. Let P_i be the point $(x_{i1},x_{i2},...,x_{ip})$ in $\mathfrak{R}_p$, representing the ith sample, whose values are given in the ith row of $\mathbf{X}$. Then x_{ik} is the marker in $\mathcal{M}$ that corresponds to the projection of P_i onto $\mathcal{M}$. Note that the marker x_{ik} is the nearest point in $\mathcal{M}$ to P_i; this property of nearness plays an important part in the following. Let $\hat{P}_i$ be the projection of P_i onto $\mathcal{L}$ and $\hat{x}_{ik}$ be the marker for the projection of $\hat{P}_i$ onto $\mathcal{M}$. The residual difference, r_i, between P_i and $\hat{P}_i$ is given by:

$$r_i^2 = \sum_{k=1}^{p}(x_{ik} - \hat{x}_{ik})^2. \qquad (7)$$

From (7), the difference $(x_{ik} - \hat{x}_{ik})$ along $\mathcal{M}$ is seen as the kth contribution to r_i^2. Using all axes, $k = 1,2,...,p$, we have that $\sum_{k=1}^{p}(x_{ik} - \hat{x}_{ik})^2$ is minimised when r_i^2 is minimised, and taking all sample points P_i ($i = 1,2,...,n$), $\|\mathbf{X} - \hat{\mathbf{X}}\|^2 = \sum_{i=1}^{n}\sum_{k=1}^{p}(x_{ik} - \hat{x}_{ik})^2$ is minimised when Σr_i^2 is minimised. As is well known (e.g. Anderson, 1958), this occurs when $\mathcal{L}$ contains G andis spanned by the principal axes of $\mathbf{X'X}$ corresponding to its r largest eigenvalues (i.e. the principal components solution). Thus, at the minimum $\hat{\mathbf{X}} = \mathbf{XV}_r$ in $\mathcal{L}$, or, relative to the original axes in $\mathfrak{R}$, $\mathbf{XV}_r\mathbf{V}_r'$. This may be written in terms of the singular value decomposition of $\mathbf{X}$ as

$$\hat{\mathbf{X}} = \mathbf{U\Sigma V'V}_r\mathbf{V}_r' = \mathbf{U\Sigma}_r\mathbf{V}_r'$$

which is the result of Eckart and Young (1936) as cited in section 1. The above has given a simple derivation of the Eckart-Young theorem and shows that $\hat{\mathbf{X}}$, obtained by projection, is the rank r approximation to $\mathbf{X}$ that minimises the residual sum-of-squares of the orthogonal projection of P_i onto $\mathcal{M}$. It is in this sense that $\hat{P}_i$ may be regarded as being an approximation to P_i for which $\hat{x}_{ik}$ predicts x_{ik}

2.4 Linear Biplot Geometry

The above has concentrated on the geometry of $\mathfrak{R}_p$ and even when discussing approximation, $\mathfrak{R}_r$, has remained embedded in the higher dimensional space. In practice, all useful interpretations must be based on information available solely in the lower-dimensional space $\mathfrak{R}_r$. In terms of the notation developed in section 2.2, the only information available is in $\mathfrak{L}$ and therefore predictions of $\hat{x}_{ik}$ by projecting onto $\mathfrak{M}$ (i.e. the original axes ξ_k) are not permissible. Plots of points $\hat{P}_i$ (i = 1,2,...,n) in $\mathfrak{L}$, approximating the total sample variation, would be greatly enhanced if one could associate values $\hat{x}_{ik}$ with each sample-point but using information contained solely in $\mathfrak{L}$.

In section 2.2 it was pointed out that $\mathfrak{N}\cap\mathfrak{L}$ contains all the points of $\mathfrak{L}$ that predict $\hat{x}_{ik}$. Because of the linearity of $\mathfrak{M}$, the normal planes $\mathfrak{N}_i$ at ξ_{ik} are parallel for all i, as are their intersections $\mathfrak{N}_i\cap\mathfrak{L}$. Thus, prediction amounts to deciding in which intersection-space each sample point of $\mathfrak{L}$ lies. What is required, is some way of associating with every sample point of $\mathfrak{L}$ at least one point, marked $\hat{x}_{ik}$, of the intersection space. This can be done quite simply by constructing any line $\mathfrak{B}$ (alias β_k) in $\mathfrak{L}$, placing the marker $\hat{x}_{ik}$ at the point where $\mathfrak{B}$ intersects $\mathfrak{N}_i\cap\mathfrak{L}$. Because of the parallelism of all these intersection spaces, the angle of intersection with $\mathfrak{B}$ is constant. Provided this angle is known, the value of the kth variable for a given sample point $\hat{P}_i$ in $\mathfrak{L}$ can be predicted by constructing the space through $\hat{P}_i$ parallel to $\mathfrak{N}\cap\mathfrak{L}$; the marker at the point where this intersects with $\mathfrak{B}$ gives the required prediction. Also, because of the parallelism, the markers on $\mathfrak{B}(\beta_k)$ relate linearly to those of $\mathfrak{M}(\xi_k)$ so that unit steps are of equal length. $\mathfrak{B}$ is a biplot axis β_k for predicting the kth variable. The geometry is shown in Figure 3.

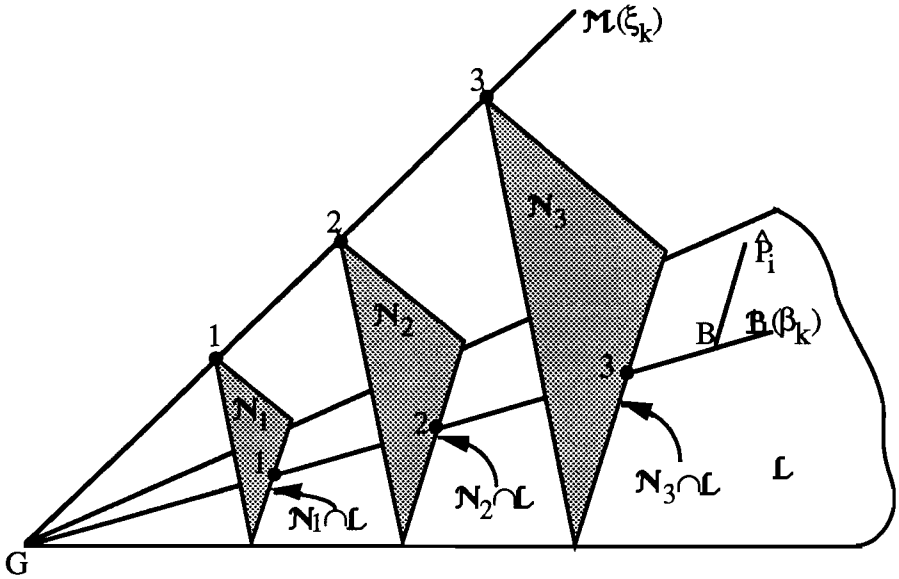

Figure 3. $\mathfrak{M}$ represents the kth Cartesian axis ξ_k, with the marked scale. $\mathfrak{B}$ is an arbitrary line in $\mathfrak{L}$ which meets the parallel spaces $\mathfrak{N}_i\cap\mathfrak{L}$ at points with markers corresponding to those of $\mathfrak{M}$. To predict $\hat{x}_{ik}$ for $\hat{P}_i$ in $\mathfrak{L}$, draw the space in $\mathfrak{L}$ that is parallel to $\mathfrak{N}_i\cap\mathfrak{L}$ meeting $\mathfrak{B}$ in B with marker $\hat{x}_{ik}$; in the diagram $\hat{x}_{ik}$ ~ 3.5. When $\mathfrak{B}$ is normal to $\mathfrak{N}_i\cap\mathfrak{L}$ then $\mathfrak{B}$ is the projection of $\mathfrak{M}$ in $\mathfrak{L}$ and the markers on $\mathfrak{B}$ are both the back-projections of the corresponding markers in $\mathfrak{M}$ and the projections of G onto $\mathfrak{N}_i\cap\mathfrak{L}$.

This construction achieves the objective of obtaining predictions from information given solely within $\mathcal{L}$ but the degree of generality is not convenient for practical use. It does, however, demonstrate the non-uniqueness of biplot axes for prediction and also shows that there is no need for $\mathcal{L}$ to be any special linear sub-space, such as that associated with principal components. It is not even necessary for $\mathcal{B}$ to pass through the origin. All one needs to know is the direction of $\mathcal{B}$ relative to $\mathcal{N}\cap\mathcal{L}$.

The whole process can be greatly simplified when (a) some definite angle of intersection is chosen in advance, and the obvious choice is a right-angle, and (b) when $\mathcal{B}$ is made to pass through the origin, G, rather than being freely located. With these choices, suppose that one of the intersection spaces $\mathcal{N}\cap\mathcal{L}$ is marked by some value $\hat{x}$, then B, the point defined by $\mathcal{B}\cap(\mathcal{N}\cap\mathcal{L})$, labels (a) the marker $\hat{x}$ for the orthogonal projection of $\hat{P}$ onto $\mathcal{B}$ and (b) the foot of the normal from G onto $\mathcal{N}\cap\mathcal{L}$. That is, B is the projection of G in $\mathcal{N}\cap\mathcal{L}$ and hence is the point in $\mathcal{N}\cap\mathcal{L}$ that is nearest the origin. Because B lies in $\mathcal{N}$, B$\hat{X}$ is normal to $\mathcal{M}$ and it follows that B is the point in $\mathcal{N}\cap\mathcal{L}$ that is nearest $\mathcal{M}$. Thus B($\hat{x}$) is the back-projection onto $\mathcal{L}$ of X($\hat{x}$) in $\mathcal{M}$ but, in this linear case, the line $\mathcal{B}$ itself is the projection of $\mathcal{M}$ onto $\mathcal{L}$. These results give further insight into the operation of biplot axes for general choices of $\mathcal{L}$. Equations (4), (5) and (6) give the matrix forms for computing B but because $\mathcal{M}$ is a line so that $\mathbf{M} = \mathbf{m}$, a column-vector, and $\mathcal{B}$ is the projection of $\mathcal{M}$ onto $\mathcal{L}$, then in this case the back-projection (6) simplifies to give:

$$\mathbf{y} = \mathbf{x}\frac{\mathbf{mm'LL'}}{\mathbf{m'LL'm}} \qquad (8)$$

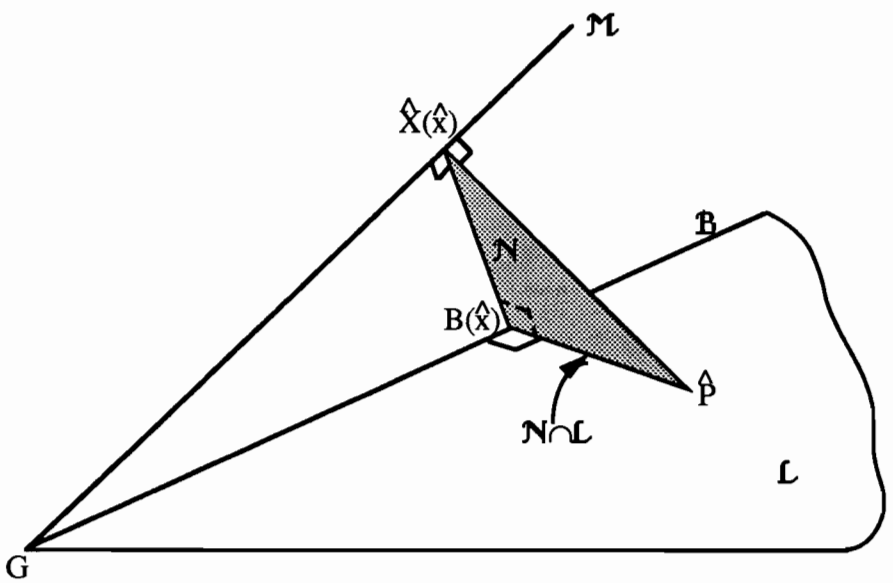

Figure 4. $\mathcal{N}$ is normal at $\hat{X}(\hat{x})$ to $\mathcal{M}$ and hence $\hat{P}$ is predicted by $\hat{X}(\hat{x})$ on $\mathcal{M}$ and also B$\hat{X}$G is a right angle. $\mathcal{B}$ is normal to $\mathcal{N}\cap\mathcal{L}$ so that $\hat{P}$ is predicted by its projection B($\hat{x}$) onto $\mathcal{B}$ within $\mathcal{L}$. Because of the normality of $\mathcal{B}$ to $\hat{P}$B, B is the nearest point in $\mathcal{N}\cap\mathcal{L}$ to G, and hence B is the nearest point of $\mathcal{N}\cap\mathcal{L}$ to $\mathcal{M}$; thus B is the back-projection of $\hat{X}(\hat{x})$ onto $\mathcal{L}$. $\mathcal{B}$ is the projection of $\mathcal{M}$ onto $\mathcal{L}$. Apart from the generality of $\mathcal{L}$, these relationships characterise the classical linear biplot.

The geometry is shown in Figure 4. The above results show that $\hat{P}$B $\perp \mathcal{B}$, $\hat{X}$B $\perp \mathcal{M}$ and $\hat{X}\hat{P} \perp \mathcal{M}$. Thus the basic requirement for predicting $\hat{X}(\hat{x})$ on $\mathcal{M}$ from $\hat{P}$ in $\mathcal{L}$ is the same as predicting B on $\mathcal{B}$ from $\hat{P}$ in $\mathcal{L}$ and then predicting $\hat{X}(\hat{x})$ on $\mathcal{M}$ from B. The latter

step is unnecessary, as B itself may be marked $\hat{x}$ uniquely on $\mathfrak{B}$, and, as required, the prediction then is accomplished entirely within $\mathfrak{L}$ and without loss of information. These orthogonality results are proved algebraically in Appendix A, where it is shown that a further consequence of back-projection is that $\hat{\mathrm{P}}\mathrm{B} \perp \hat{\mathrm{X}}\mathrm{B}$.

None of the results discussed in this section depend on the choice of $\mathfrak{L}$. Biplot axes, both for prediction and interpolation, are defined for any sub-space $\mathfrak{L}$; this degree of generality should be no surprise, as coordinate axes are not dependent on the statistical quality of the configurations that they are used to represent. When $\mathfrak{L}$ is a best-fitting linear sub-space, as with components analysis, we arrive at the classical biplot, but nothing else changes.

2.5 Prediction and Interpolation

Section 2.4 discussed the geometry of predicting $\hat{x}$ for a point $\hat{\mathrm{P}}$. It was shown that for given $\mathfrak{L}$ the biplot axes are not uniquely defined for prediction, although there is a unique definition that is of special practical significance. In section 1 interpolation was expressed as the vector-sum $\sum_{k=1}^{p} x_k \mathbf{e}_k$. With orthogonal axes in $\mathfrak{R}_p$ this gives the same interpolated and predicted values, i.e. at the uniquely defined point P at the intersection of the normal planes $\mathfrak{N}_k$ at x_k on ξ_k ($k = 1,2,...,p$). The same method does not work with biplot axes β_k ($k = 1,2,...,p$) because in $\mathfrak{L}$, of $r < p$ dimensions, the p normal planes usually will not intersect in a common point. What is required for interpolation is the position $\hat{\mathrm{P}}$ of the projection in $\mathfrak{L}$ of P, which is given by:

$$(\sum_{k=1}^{p} x_k \mathbf{e}_k)\mathbf{L}\mathbf{L}' = \sum_{k=1}^{p} x_k (\mathbf{e}_k \mathbf{L}\mathbf{L}'). \qquad (9)$$

Thus $\mathbf{e}_k\mathbf{L}\mathbf{L}'$ corresponds to one unit on the biplot axis β_k . This axis is the same as the previous biplot axis for prediction because it remains the projection of $\mathfrak{M}$ onto $\mathfrak{L}$. However $\mathbf{e}_k\mathbf{L}\mathbf{L}'$ represents an orthogonal projection of $\mathbf{e}_k$ and not its back-projection (8) as was required for prediction. With the understanding that different scales are required, the same biplot axes can be used both for prediction and for interpolation, the latter being determined by the vector-sum (9). The biplot axes for interpolation are uniquely determined and the variant axes described in section 2.4 that are acceptable for prediction cannot be used for interpolation, thus providing another reason for choosing β_k to be the unique projections of ξ_k ($k = 1,2,...,p$) onto $\mathfrak{L}$ which serves for both purposes.

Thus, in the linear case, the two sets of axes can be made to coincide and, apart from the minor inconvenience of requiring different scales, the simplicity of having the same axes both for interpolation and prediction adds to the simplicity of linearity. To emphasise the differences may be seen as unnecessary mathematical obfuscation but in the following sections it will be seen that without linearity, different sets of axes are needed for interpolation and prediction. Linearity masks what become essential distinctions in non-linear extensions but to understand the generalisations which follow, it is essential first to have a clear understanding of the permissible generalities of the linear case.

2.6 Variant Linear Biplots

The remarks of section 2.4 have made it evident that any line β_k in $\mathbf{L}$ suffices for prediction of ξ_k. Because it is any line, β_k, which will now be referred to as β, must suffice for the prediction of all variables ξ_k (k = 1,2,...,p). Thus only one biplot axis is needed for all predictions. There are two problems with this apparent simplification. Firstly β would have to be marked with p scales, which would be unreadable. Secondly, prediction for a point in $\mathbf{L}$ of r dimensions would involve constructing $(r - 1)$-dimensional planes parallel to the different directions $\mathbf{N}_k$; the latter is impracticable unless $r = 2$. When $r = 2$, the intersection spaces $\mathbf{N}_k \cap \mathbf{L}$ become lines, and it is necessary to record their directions in $\mathbf{L}$ as well as their scales; a star-like icon whose rays give direction and scale for each variable, is one possibility. For two axes, ξ_h and ξ_k, say, one may choose β to pass through the intersections of equal markers on $\mathbf{N}_h \cap \mathbf{L}$ and $\mathbf{N}_k \cap \mathbf{L}$; then both scales on β would coincide; such a biplot axis will named β_{hk}. The situation is shown Figure 5.

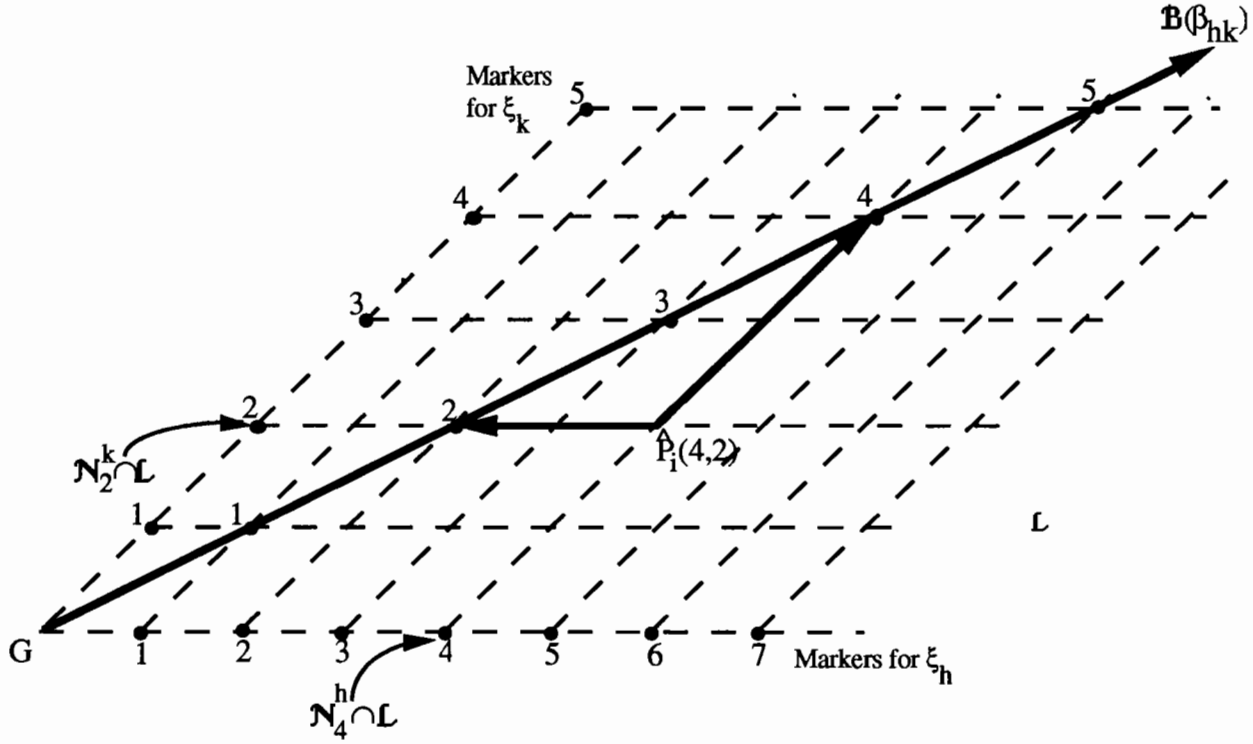

Figure 5. A single biplot axis β_{hk}, for predicting variables ξ_h and ξ_k. The parallel lines are the intersection sets $\mathbf{N}_i^h \cap \mathbf{L}$ and $\mathbf{N}_i^k \cap \mathbf{L}$. Predictions for both variables are made on β_{hk}, as shown.

To predict, one needs to know the two parallel directions and then move parallel to the direction for ξ_k to predict $\hat{x}_h$ and parallel to ξ_h to predict $\hat{x}_k$. This operation is demonstrated in Figure 5; note the same device can be used with conventional x-y plots. Clearly, in r-dimensional biplots the same principle can be used to derive axes which simultaneously predict r variables using a single scale. Such devices reduce the number of biplot axes needed for prediction to the smallest integer greater than p/r and they allow the rankings of two variables to be compared fairly readily; it remains to be seen if such multiple-scale axes have any real practical utility.

The results of this section are valid for any $\mathbf{L}$, including approximation spaces that have been chosen to optimise some property of the samples - usually concerned with inter-

sample distance. The axes β or β_{hk} are not normally back-projections of any axis ξ_k. There remains the possibility of choosing $\mathbf{L}$ to optimise some property of the biplot axes. For example it is shown in appendix B that $\mathbf{L}$ may be chosen so that any p-1 axes ξ_k may be represented by back-projections to give a single biplot axis in two dimensions. Of course, such a choice of $\mathbf{L}$ is unlikely to give good approximations to sample variation.

3 Non-linear Biplots

Gower and Harding (1988) introduced the notion of a non-linear biplot for quantitative variables; the linear biplot is a special case. Although it is convenient to represent samples relative to orthogonal axes, there is no obligation to do so. An alternative approach is to define a function $d_{ij} = f(\mathbf{x}_i,\mathbf{x}_j)$ giving the distance between samples i and j ($i,j = 1,2,...,n$). With linear biplots, d_{ij} is Pythagorean distance, defined by $d_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)'$.

Using the notation $\mathbf{A} = \{a_{ij}\}$ here, and throughout, to denote a matrix $\mathbf{A}$ with general term a_{ij}, we shall require for formal mathematical reasons that:

(i) the $n \times n$ matrix $\{d_{ij}\}$ shall be symmetric, $d_{ij} \geq 0$ and $d_{ii} = 0$ for all i,j

(ii) $\{d_{ij}\}$ is embeddable in Euclidean space.

Requirement (ii) implies that n points can be found in $\mathfrak{R}_{n-1}$ that reproduce the distances d_{ij}. In practice, (ii) may be relaxed, provided the departures from Euclideanarity are not "too serious".

Thus a configuration of sample-points P_i ($i = 1,2,...,n$) in $\mathfrak{R}_{n-1}$ is given, with the distance d_{ij} between points P_i and P_j. The terminology will be used that the points P_i ($i = 1,2,...,n$) generate the distances d_{ij}. The dimensionality of the full space is now n -1 rather than the previous value of p. The configuration has arbitrary axes and the problem arises of associating with the sample-points, information on the original variables. This will be done by finding appropriate non-linear coordinate axes. The configuration of samples in the full space may be approximated in r dimensions and the biplot problem than becomes that of describing what forms are taken by the non-linear axes in this approximation and what are their properties.

In the linear case, the position X of any sample is fixed when its Pythagorean distances from all the points P_i ($i = 1,2,...,n$) are known. Suppose now that X is defined by sample values ξe_k, then the locus of X as ξ varies reproduces ξ_k, the kth Cartesian axis. Thus the conventional process has been inverted and rather than the axes determining the points, now the points determine the axes. When a general distance d_{ij} is used, this process gives a curvilinear locus of X, termed a trajectory. There is one trajectory for each of the p variables and these trajectories are concurrent, at a point O, say, because $\xi = 0$ is common to them all. Because $\mathbf{X}$ is centred at its mean, the point O corresponds to a sample taking the mean values of all the original variables. The method has very general applicability but here the special case is followed for which Gower and Harding (1988) gave a detailed algebraic exposition. It is assumed that:

$$\text{(a) } d_{ij}^2 = \sum_{k=1}^{p} f_k(x_{ik}, x_{jk}) \qquad (10)$$

(b) The points P_i (i = 1,2,...,n) are derived from $\mathbf{D} = \{-\frac{1}{2}d_{ij}^2\}$ by principal coordinates analysis (Gower, 1966), alias classical scaling (Torgerson, 1955).

In (a) it is assumed that each variable contributes independently to total squared-distance. This may seem to be a restrictive assumption but nevertheless it includes many distance-functions in common use (see e.g. Gower and Legendre, 1986). If it is further assumed that $f_k(x_{ik}, x_{jk}) = 0$ iff $x_{ik} = x_{jk}$, it easily follows that the functions for the individual variables themselves define distances. Defining $\mathbf{D}_k = \{-\frac{1}{2}f_k(x_{ik}, x_{jk})\}$ where k is fixed, (10) may be written:

$$\mathbf{D} = \sum_{k=1}^{p} \mathbf{D}_k. \tag{11}$$

Assumption (b) implies that coordinates of P_i (i = 1,2,...,n) can be found in, at most, n-1 dimensions as the rows of the $n \times (n\text{-}1)$ matrix $\mathbf{Y}$, where $\mathbf{\Lambda} = \mathbf{Y}'\mathbf{Y}$ is diagonal, because $\mathbf{Y}$ is referred to its principal axes. In this representation $\mathbf{e}'\mathbf{Y} = \mathbf{0}$, so the centroid of the sample-points P_i (i = 1,2,...,n) is at the origin G, which usually differs from O, the concurrent point of the trajectories.

A coordinate representation in $\mathfrak{R}_n$ for any new point $\boldsymbol{\xi} = (\xi_1,\xi_2,...,\xi_p)$ may be found by using a result, Gower (1968), which shows that if $f(\boldsymbol{\xi},\mathbf{x}_i)$ gives the squared distance of $\boldsymbol{\xi}$ from P_i (i = 1,2,...,n) then the first n -1 dimensions of the coordinates of $\boldsymbol{\xi}$ are given by:

$$\mathbf{z} = -\tfrac{1}{2}\mathbf{g}'\mathbf{Y}\mathbf{\Lambda}^{-1} \tag{12}$$

where $\mathbf{g} = \mathbf{f} + 2\mathbf{D}\mathbf{e}/n$ and $\mathbf{f} = \{f(\boldsymbol{\xi},\mathbf{x}_i)\}$. One further dimension is required to represent $\boldsymbol{\xi}$ exactly. Its coordinate z, in the nth dimension, is easily calculated as:

$$z^2 = d_\xi^2 - \tfrac{1}{4}\mathbf{g}'\mathbf{Y}\mathbf{\Lambda}^{-2}\mathbf{Y}'\mathbf{g} \tag{13}$$

where d_ξ represents the distance of $\boldsymbol{\xi}$ from G. To find d_ξ requires the result that the squared distance between the centroids of two sets of points given in a partitioned $(n + m)\times(n + m)$ squared-distance matrix $\begin{pmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}' & \mathbf{B} \end{pmatrix}$, including the factor $-\frac{1}{2}$, is given by:

$$\frac{1}{n^2}\mathbf{e}'\mathbf{A}\mathbf{e} + \frac{1}{m^2}\mathbf{e}'\mathbf{B}\mathbf{e} - \frac{2}{nm}\mathbf{e}'\mathbf{C}\mathbf{e}. \tag{14}$$

The $(n + 1)\times(n + 1)$ squared-distance matrix for the n original samples augmented by $\boldsymbol{\xi}$ is:

$$\begin{pmatrix} \mathbf{D} & -\frac{1}{2}\mathbf{f} \\ -\frac{1}{2}\mathbf{f}' & 0 \end{pmatrix}, \tag{15}$$

and applying (14) shows that :

$$d_\xi^2 = \frac{\mathbf{e}'\mathbf{D}\mathbf{e}}{n^2} + \frac{\mathbf{e}'\mathbf{f}}{n} \tag{16}$$

which allows z to be calculated from (13). Thus, appending (13) to (12) gives the full set $(\mathbf{z},z)$ of coordinates in $\mathfrak{R}_n$ of $\boldsymbol{\xi}$, relative to the coordinate system of $\mathbf{Y}$.

In general, every point that is added introduces an extra dimension given by (16). Fortunately this is not the problem that it might seem, when $\mathbf{L}$ is defined, as below, as the space spanned by some, or all, of the columns of $\mathbf{Y}$. Firstly, all the extra dimensions are then orthogonal to $\mathbf{L}$, so have no effect on projections onto $\mathbf{L}$. Secondly, the mutual orthogonality of the extra dimensions, guarantees that they are irrelevant for defining tangents to the trajectories (section 3.3). Thirdly, when they satisfy the relationship $\mathbf{q}\mathbf{L} = 0$

derived from (A8) of appendix A, one may proceed as if the extra dimensions do not exist. These three properties imply that the extra dimensions can often be ignored but they have to be allowed for when they contribute to distances, especially when defining neighbour-regions with categorical variables (section 4.3).

3.1 Non-linear biplots - Full Space

With these preliminaries, Gower and Harding (1988) defined for the kth variable a pseudo-sample $\xi\mathbf{e}_k$ and established that its squared distances from the original n samples are given by:

$$\mathbf{f} = \{f(\xi,x_{ik})\} - \{f(0,x_{ik})\} + \sum_{k=1}^{p}\{f(0,x_{ik})\}, \qquad (17)$$

which may be immediately substituted into (12), (13) and (16) to give coordinates for the pseudo-sample.

The only term in (17) that involves ξ is $f(\xi,x_{ik})$. This has the fundamental consequence that the point ξ on the kth trajectory is nearest P_i when $f(\xi,x_{ik})$ is minimal, which has the simple solution $\xi = x_{ik}$. This important result shows that the kth trajectory is acting like conventional coordinate axes, at least insofar as prediction of the coordinates for a sample-point P_i is given by the nearest markers on the p trajectories. In other words the coordinates are given by dropping normals from P_i onto the trajectories, the equivalent of orthogonal projection, and reading off the marker at the point of intersection. With curvilinear trajectories, there may be several normals from P_i to each trajectory but it is clear that it is the shortest normals that are required and, except in pathological cases, these will be uniquely defined.

If a sample with values $(x_{i1},x_{i2},...,x_{ip})$, is located by constructing each of the $(n$-2)-dimensional linear spaces orthogonal to the kth trajectory at x_{ik} $(k = 1,2,...,p)$, these p spaces will intersect, in n-p-1, or more, dimensions. Thus not only in the approximation space but now also in the full space, a whole region predicts the same values. However, each sample-point of a region predicts the unique values given by the nearest marker on each trajectory. The p trajectories continue to act like conventional linear coordinate axes.

3.2 Non-linear Biplots - Interpolation

Next we examine what happens in r-dimensional approximations. As explained above, the assumption (b) is the basis of the approximation examined here, and this implies that P_i is approximated by $\hat{P}_i$ which is the orthogonal projection of P_i onto some r-dimensional subspace $\mathbf{L}$. The trajectories themselves may also be projected onto $\mathbf{L}$ in a similar manner to the way that orthogonal Cartesian coordinates are projected onto $\mathbf{L}$ to obtain linear biplot axes. When $\mathbf{L}$ is given by principal coordinates, it follows that the first r columns of (12) give the coordinates of $\boldsymbol{\xi}$ projected onto $\mathbf{L}$; the extra dimensions given by (13) are orthogonal to $\mathbf{L}$ so are immaterial. In particular any point $\hat{\xi}$ on the kth trajectory can be projected onto $\mathbf{L}$, and thence the whole trajectory, which may be termed the kth biplot trajectory β_k.

Suppose now that $\boldsymbol{\xi} = (\xi_1,\xi_2,...,\xi_p)$ is a point to be interpolated into $\mathcal{L}$ by (12). Because of the independence assumption (a), $\mathbf{g}$ may be written:

$$\mathbf{g} = \sum_{k=1}^{p}\left(\mathbf{f}_k + 2\mathbf{D}_k\mathbf{e}/n\right)$$

where $\mathbf{f}_k$ is given by (17) with ξ replaced by ξ_k, so that (12) becomes:

$$-\tfrac{1}{2}\sum_{k=1}^{p}\left(\mathbf{f}_k + 2\mathbf{D}_k\mathbf{e}/n\right)'\mathbf{Y}_r\boldsymbol{\Lambda}_r^{-1}$$

where now $\mathbf{Y}_r$ represents the first r columns of $\mathbf{Y}$, which determine $\mathcal{L}$, and similarly for $\boldsymbol{\Lambda}_r$. The kth term in the summation is the projection onto $\mathcal{L}$ of the point marked ξ_k on the kth trajectory, and hence corresponds to the marker ξ_k on the kth biplot trajectory for interpolation. Thus for interpolation, the vector-sum method remains valid for non-linear biplot trajectories.

3.3 Non-linear Biplots - Prediction

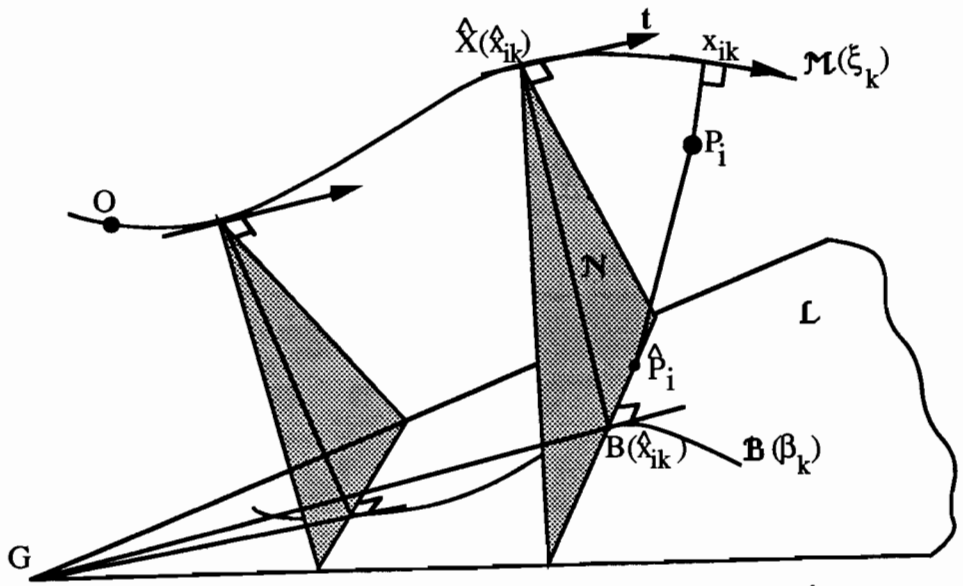

Figure 6. P_i has coordinate (marker) x_{ik} on $\mathcal{M}$ and is approximated by $\hat{P}_i$ in $\mathcal{L}$, which has coordinate $\hat{X}(\hat{x}_{ik})$ on $\mathcal{M}$. $\mathcal{M}$ maps $\mathcal{B}$ in $\mathcal{L}$ as the locus of the projections of G onto $\mathcal{N}\cap\mathcal{L}$ where $\mathcal{N}$ is normal to the tangents to $\mathcal{M}$. Thus the coordinate predicted for $\hat{P}_i$ in $\mathcal{L}$ is $B(\hat{x}_{ik})$, obtained by subtending right-angles from $G\hat{P}_i$ to $\mathcal{B}$. $\mathcal{M}$ need not meet $\mathcal{L}$. The non-parallelism of the normal-planes is indicated.

Now consider prediction. The geometry is shown in Figure 6, the non-linear counterpart of Figure 3. $\hat{P}_i$ is predicted by $\hat{x}_{ik}$ on the kth trajectory by dropping the (shortest) normal from $\hat{P}_i$. The linear space $\mathcal{N}$, orthogonal to the kth trajectory at $\hat{x}_{ik}$ contains all points to be predicted by $\hat{x}_{ik}$ and hence contains $\hat{P}_i$ itself. Within $\mathcal{L}$, all points in $\mathcal{N}\cap\mathcal{L}$ are to be predicted by $\hat{x}_{ik}$ and the problem remains of finding a biplot trajectory to which all points in $\mathcal{N}\cap\mathcal{L}$ can be uniquely associated. In the non-linear case, the orthogonal projection interpolation biplot trajectories, discussed in 3.2, are not acceptable, because as $\hat{x}_{ik}$ varies the resulting normal spaces $\mathcal{N}_k$ are not parallel as they are in the linear case; hence there is no constant reference angle within $\mathcal{L}$.

The problem of obtaining a suitable prediction biplot trajectory in $\mathcal{L}$ may be linearised by approximating the kth trajectory $\mathcal{M}$ at ξ by a linear segment, essentially the tangent $\mathbf{t}$ at ξ. This determines $\mathcal{M}$ locally and, by the arguments of section 2.4, the corresponding point on the prediction biplot trajectory is given by some point B in $\mathcal{N} \cap \mathcal{L}$. As is shown in appendix A the back-projection that is nearest the axis ξ will no longer be convenient, because the offset of $\mathbf{t}$ induces an oblique angle between its image in $\mathcal{L}$ and $\mathcal{N} \cap \mathcal{L}$. Further this oblique angle will vary with $\mathbf{t}$. Appendix A shows that the orthogonal projection of G onto $\mathcal{N} \cap \mathcal{L}$, that is the point in $\mathcal{N} \cap \mathcal{L}$ that is nearest G, is still given by (6). It is this point that should be labelled B and marked ξ. All points in $\mathcal{N} \cap \mathcal{L}$ will make a right-angle with the line GB. As ξ moves along the kth trajectory in the full space, the locus of B will trace out β_k the kth biplot prediction trajectory. To use this form of prediction-biplot for a point $\hat{P}_i$ in $\mathcal{L}$, it is only necessary to read off the marker at the point where $G\hat{P}_i$ subtends a right-angle on the trajectory.

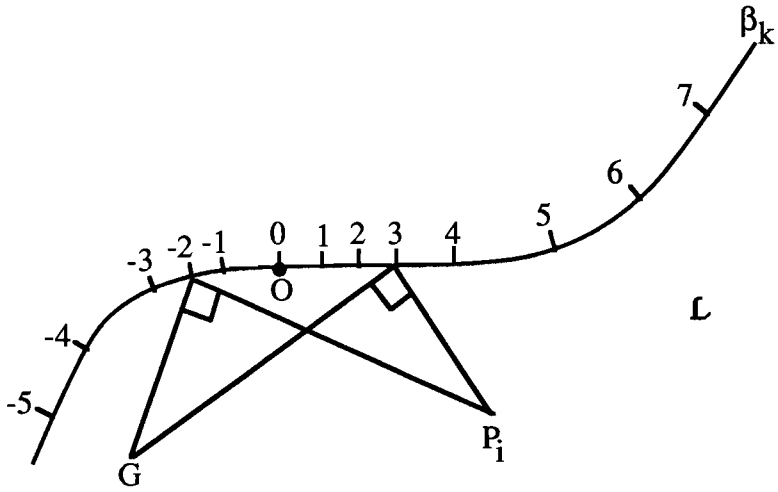

Figure 7. Non-linear biplot axis, showing a non-uniform scale. The point P_i has $b_{ik} = 3$ for its predicted value. A second construction gives $b_{ik} = -2$ but this is not the nearest to P_i, so is excluded. O is the point common to all trajectories and G the centroid.

The trajectories can be computed almost as simply as can be the interpolation trajectories. Equations (12), (13) and (16) give the parametric form of the kth trajectory. As described at the end of section 3, the extra dimensions given by (13) and (16) are irrelevant, so the tangent at any point is easily obtained by differentiation of (12) in the usual way; the exact form will depend on the choice of distance function. This tangent determines $\mathcal{M}$ and thence equation (6) of the linear case may be used to find B in $\mathcal{L}$. This calculation has to be repeated as ξ moves along the kth trajectory, $\mathcal{M}$, changing with the changing tangents. Thus, prediction of the value of the kth variable for a point P_i in $\mathcal{L}$ is merely a matter of subtending orthogonally onto β_k, the corresponding non-linear axis, and noting the value of the marker, as is shown in Figure 7.

With a linear biplot, subtending a right angle is the same as orthogonal projection onto β_k. Again the non-linear case is seen to be similar to the linear case and uses only information given in the approximation space. The major difference from the linear case is that not only do the interpolation and prediction scales differ but so do the trajectories themselves. Thus

it is no longer possible to have a single biplot diagram that can be used both for interpolation and prediction; two quite separate diagrams must be used.

As with linear biplots, non-linear prediction trajectories are not uniquely defined and any trajectory in $\mathbf{L}$ will suffice, provided there is a rule that allows the whole of $\mathbf{N} \cap \mathbf{L}$ to be constructed at every point of the trajectory. Subtending right-angles, described above, is one such rule. Another is to construct the trajectory in $\mathbf{L}$ for which $\mathbf{N} \cap \mathbf{L}$ is given by the normal to that trajectory. I have not examined the details of this construction. In the linear case, both rules give the same trajectory but in the non-linear case they differ.

4. Generalised Biplots and Categorical Variables

Gower (1991) shows how non-linear biplots can be modified to give generalised biplots that allow the inclusion of categorical variables. Non-linear biplots, and hence linear biplots, are essentially special cases. Categorical variables take a finite number, l_k, of values, often termed levels. The term *value* will be used to emphasise that x_{ik} may represent a numerical value for a kth continuous variable or a category-level for a kth categorical variable. Thus if the kth variable is colour, then x_{ik} may have a value Blue or Green or Red, (say), in which case $l_k = 3$. Because categories are not continuous, they cannot be represented relative to continuous coordinate axes. However, they may be represented by a finite number of points, referred to as the category-level points (CLPs), or as we shall see, by regions labelled with the category values, such as Blue, Green, Red.

In generalised biplots, the position of a marker ξ for the kth variable is obtained as follows. Consider the set of pseudo-samples $\mathbf{x}_i + (\xi - x_{ik})\mathbf{e}_k$ (i = 1,2,...,n) and find their positions relative to the axes of $\mathbf{Y}$ by (12). The centroid of these n points is taken as the point to be associated with the marker ξ for the kth variable. As ξ varies this centroid traces out a trajectory for continuous variables and for categorical variables, determines l_k distinct points, one for each value of ξ that corresponds to a category-level. Unfortunately, for continuous variables $\xi = 0$ is not a common marker on all trajectories, which therefore are usually non-concurrent. In section 4.2 it will be shown that this defect can be overcome for interpolation but not for prediction.

4.1 Generalised Biplots - Full Space

Gower (1991) shows that all the squared distances for a fixed ξ of the pseudo-sample may be accumulated into a matrix (compare equation (15)):

$$\begin{pmatrix} \mathbf{D} & -\frac{1}{2}\mathbf{F} \\ -\frac{1}{2}\mathbf{F}' & \mathbf{D} - \mathbf{D}_k \end{pmatrix} \qquad (18)$$

where the first n rows (columns) refer to the n original samples and the second n rows (columns) to the pseudo-samples. In (18), $\mathbf{F} = -2\mathbf{D} + 2\mathbf{D}_k + \mathbf{f}_k\mathbf{e}'$ where $\mathbf{f}_k = \{f(\xi, x_{ik})\}$.

From (14) the squared distances from the centroid of the pseudo-samples to each of the n original samples are given by the rows of:

$$\left(\frac{\mathbf{e}'(\mathbf{D} - \mathbf{D}_k)\mathbf{e}}{n}\mathbf{I} - 2(\mathbf{D} - \mathbf{D}_k)\right)\frac{\mathbf{e}}{n} + \mathbf{f}_k \qquad (19)$$

which gives the vector $\mathbf{f} = \{f(\xi,\mathbf{x}_i)\}$ to be used with (12). The only term dependent on ξ is $\mathbf{f}_k$. Thus the nearest point on the kth axis to the ith sample is when f_{ik} is zero, i.e. $f(\xi,x_{ik}) = 0$ or $\xi = x_{ik}$. This is true for both quantitative and categorical variables. Thus the vital nearness property of non-linear and linear axes remains true for these generalised axes. Because $\mathbf{e}'\mathbf{Y} = \mathbf{0}$, the constant term $\frac{(\mathbf{e}'(\mathbf{D} - \mathbf{D}_k)\mathbf{e})}{n^2}\mathbf{e}$ makes no contribution to (12) and the form of (19) admits the further simplification:

$$\mathbf{z}_k = -\tfrac{1}{2}(\mathbf{f}_k + 2\mathbf{D}_k\tfrac{\mathbf{e}}{n})'\mathbf{Y}\mathbf{\Lambda}^{-1}. \qquad (20)$$

Gower (1991) shows that, for generalised biplot axes, the n coordinates corresponding to the actual values observed for the kth variable are obtained from (12) as:

$$\mathbf{Z} = (\mathbf{I} - \mathbf{e}\mathbf{e}'/n)\mathbf{D}_k\mathbf{Y}\mathbf{\Lambda}^{-1}. \qquad (21)$$

For categorical variables, (21) has only l_k different rows corresponding to the distinct CLPs, the remaining rows being superfluous repetitions. For categorical variables, (21) gives all the CLPs but for continuous variables it gives n points, which may not be enough, or well-enough distributed, points to ensure a smooth trajectory. Then, it may be necessary to augment (21) by calculating extra points on the trajectory corresponding to unobserved values of ξ in (17) and (12). As has already been observed, for continuous variables the trajectories are not concurrent; an adjustment can be made to ensure concurrency when interpolating (section 4.2) but this seems not possible for prediction. Otherwise, for continuous variables the same geometry applies as described in section 3 for non-linear biplots and no further comment is required.

For categorical variables, the extra dimensions induced by the coordinates of (13) and (16) now require that dimensionality n^* is given by $n^* = n + l_k - 1$. It has been shown that in the full space $\mathfrak{R}_n{}^*$, every sample is nearest its CLP. Thus in $\mathfrak{R}_n{}^*$, if we wish to predict (section 4.3) the categorical level of a sample, it is sufficient to find the nearest CLP. Thus for every categorical variable, $\mathfrak{R}_n{}^*$ may be partitioned into neighbour-regions, each containing all the samples with a particular category-level and no others.

It follows immediately from (21) that $\mathbf{e}'\mathbf{Z} = \mathbf{0}$ so that for each categorical variable the mean of its CLPs weighted by the number of occurrences of each level, is at the origin G, the centroid of the ordination. This result does not apply to the extra dimensions induced by (13), showing that the subspace $\mathfrak{M}$ of CLPs is disjoint from the space spanned by the columns of $\mathbf{Y}$, from which $\mathfrak{L}$ is derived. Nevertheless it can be expected that the two spaces are close in the vicinity of G. The offset $\mathbf{q}$ from G lies only partly in the space of the extra dimensions, so $\mathbf{q}\mathbf{Y} \neq \mathbf{0}$, and in approximations, $\mathbf{q}\mathbf{L} \neq \mathbf{0}$, the condition derived in appendix A for the back-projection formula (6) to be valid for non-intersecting spaces $\mathfrak{L}$ and $\mathfrak{M}$. It follows that the full formulae (4) and (5) have to be used for back-projection

4.2 Generalised Biplots - Interpolation

In the approximation space $\mathfrak{L}$, interpolation by vector-sums remains valid with the trajectories and CLPs of $\mathfrak{R}_{n-1}$ projected orthogonally onto $\mathfrak{L}$. Again, as the coordinate z

given by (13) is orthogonal to $\mathfrak{L}$, the first r columns of (21) give the CLPs in $\mathfrak{L}$. However, the generalised biplot axes for continuous variables, as defined above, are not concurrent. This is an inconvenience that may be avoided by noting that, so long as they sum to zero, any vectors may be added to each $\mathbf{z}_k$ of (20) without affecting interpolation. Writing O_k to denote the point on the kth trajectory that corresponds to the mean of the kth quantitative variable, and $\bar{O}$ to denote the centroid of all the O_k, all that has to be done to ensure that the vectors are concurrent at $\bar{O}$, is to subtract from $\mathbf{z}_k$ the vector corresponding to $O_k - \bar{O}$. The position of the CLPs for categorical variables need not be changed. With this adjustment, Figure 8 shows interpolation with a mixture of quantitative and categorical variables.

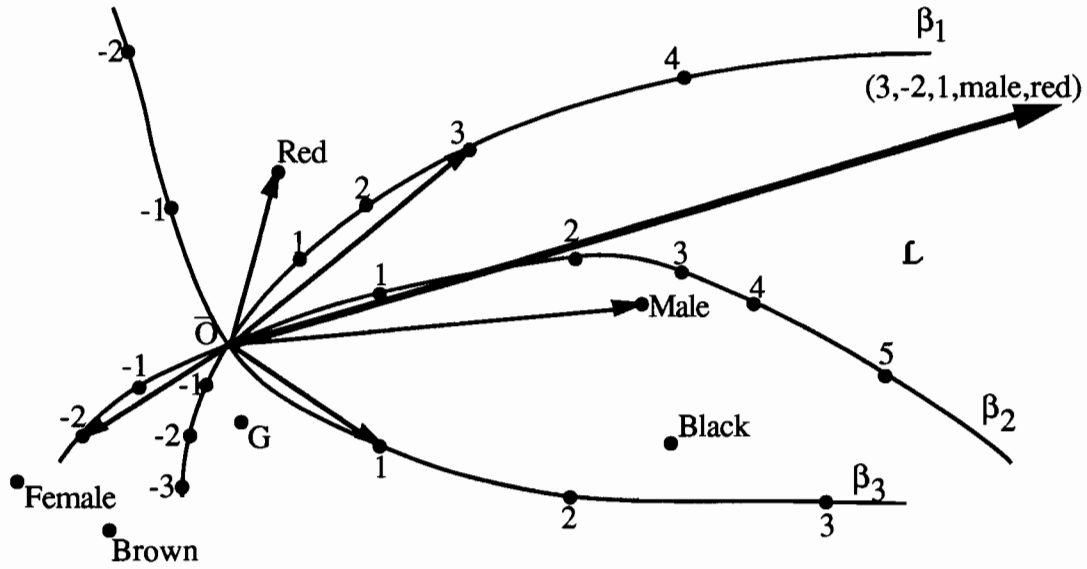

Figure 8. Generalised biplot. Interpolation as a vector-sum of a sample with three quantitative variables and two categorical variables.

4.3 Generalised Biplots - Prediction (Categorical Variables)

Just as prediction trajectories differ from interpolation trajectories for continuous variables, so do the CLPs for interpolating with, and predicting for, category values. Consider again the categorical variable Colour, with three values Red, Green and Blue. These are represented by three points in $\mathfrak{R}_n^*$ which determine a triangle in a two-dimensional space $\mathfrak{M}$. In general, the CLPs form an $s = (l_k - 1)$-dimensional simplex in $\mathfrak{M}$ and this is the reason that the result of section 2.2 was required in its general form. Because of the nearness property, the space $\mathfrak{R}_n^*$ comprising $\mathfrak{M}$ and its normal-space $\mathfrak{N}$, may be partitioned into neighbour-regions each of which contains all the samples with a particular category-level. The intersections of these neighbour-regions with $\mathfrak{L}$ give all points in $\mathfrak{L}$ that are predicted by the corresponding category-levels. This intersection is given by the back-projection of the neighbour-regions of $\mathfrak{M}$, and where necessary of $\mathfrak{N}$, onto $\mathfrak{L}$. Samples which fall in the back-projected neighbour-regions of $\mathfrak{L}$ are predicted to have the corresponding labelled category-levels. Inevitably, the ordinary orthogonal projections of the samples used to give $\mathbf{Y}_r$ in principal coordinates analysis will mean that some sample-points will occupy wrongly labelled prediction-regions of $\mathfrak{L}$. Thus with categorical variables, the distinction between interpolation and prediction is even more

marked than it is for continuous variables, for now interpolation uses sets of points, the projected CLPs, while prediction uses whole regions of space. The details of the construction of prediction-regions are technically quite intricate and to avoid an imbalance in this presentation, only an outline is given here of the problems involved; a much fuller discussion is given by (Gower, 1992).

In general each of $\mathfrak{L}$, $\mathfrak{M}$ and $\mathfrak{N}$ may be of two parts. One part of $\mathfrak{L}$ consists of points that can be formed from the back-projection of a part $\mathfrak{M}^*$ of $\mathfrak{M}$; the other part of $\mathfrak{L}$ can be formed from back-projection of a part $\mathfrak{N}^*$ of $\mathfrak{N}$. $\mathfrak{M}^*$ and $\mathfrak{N}^*$ are, respectively, the orthogonal projections of $\mathfrak{L}$ onto $\mathfrak{M}$ and $\mathfrak{N}$. There remain parts of $\mathfrak{M}$ and $\mathfrak{N}$ that do not back-project onto $\mathfrak{L}$ and so play no part in the formation of the prediction-regions. In practice, many of these subspaces may be null. Much depends on whether $r \geq s$ or $r < s$, where now s is to be interpreted mostly as the dimension of $\mathfrak{M}^*$ but also sometimes of $\mathfrak{N}^*$. When $r \geq s$, equations (4), (5), but not (6), allow the calculation of the back-projections into $\mathfrak{L}$ of any point in $\mathfrak{M}^*$; note that (6) is unavailable because z, given by (13), does not satisfy the condition $\mathbf{qL} = \mathbf{0}$ of appendix A is satisfied.

A case of special practical importance is when $l_k = 3$ ($s = 2$) and $r = 2$. Then one need back-project only 4 points, the 3 vertices of the simplex and its circumcentre, which it is convenient to label with the name of the categorical variable itself. This information is all that is needed to construct the back-projections of the neighbour regions in $\mathfrak{L}$. Back-projection transforms mid-points into mid-points, so the boundaries of the prediction regions can be constructed but, because nearness properties are not preserved under back-projection, they are not neighbour-regions in $\mathfrak{L}$. Whenever $r = s$ neighbour-regions can be similarly characterised but when $r \geq 3$, practical uses are limited.

When $s = l_k - 1 > r$ things become more complicated. The neighbour-regions in $\mathfrak{M}$ are fairly straightforward but those within $\mathfrak{M}^*$ are not. When $r = 2$, this situation will arise whenever there are categorical variables with four, or more, levels - as is not unusual. What has to be done, is:

(1) Project $\mathfrak{L}$ onto $\mathfrak{M}$ to give vectors whose columns span that part $\mathfrak{M}^*$ of $\mathfrak{M}$ which back-projects onto $\mathfrak{L}$.
(2) Examine the intersection of $\mathfrak{M}^*$ with the neighbour-regions of the simplex defined by the category-level points.
(3) Characterise these intersections by a minimal set of points.
(4) Back-project the minimal set onto $\mathfrak{L}$ to give the final configuration for the neighbour-regions in $\mathfrak{L}$.
(5) When $\mathfrak{N}^*$ is not null, the back-projection of step (3) must be augmented by orthogonal extension (see appendix A) into the corresponding part of $\mathfrak{L}$.

The above has indicated the kinds of problem that have to be solved for constructing prediction-regions for categorical variables (Gower, 1992). Alternatively, the back-projection of the CLPs give a simple approximate method that might suffice for doing most predictions by eye. Then all that would be seen in a plot of $\mathfrak{L}$ are points for Blue, Red and Green, and similarly for other categorical variables. This approximation, when adequate, has some attraction as it overcomes the impracticability of exhibiting prediction-regions for

all categorical variables simultaneously. Unfortunately, projection does not preserve nearness, so this approximation could be misleading if the nearness of a sample-point to a projected CLP were used for interpretation - and it is difficult to see what other tool might be available. For safety, one needs to exhibit prediction-regions in $\mathcal{L}$ for each variable separately, together with back-projected, not orthogonally projected, samples.

5. Relationship of Non-linear to Generalised Biplots

The key results for distinguishing the two methods are (17) and (19), which give the distances of the the pseudo-samples from the points P_i. For comparison they are presented again, with (19) in a slightly rearranged form:

$$\{f(\xi,x_{ik})\} - \{f(0,x_{ik})\} + \sum_{k=1}^{p}\{f(0,x_{ik})\},$$

$$\{f(\xi,x_{ik})\} - \left(\frac{\mathbf{e}'\mathbf{D}_k\mathbf{e}}{n}\mathbf{I} - 2\mathbf{D}_k\right)\frac{\mathbf{e}}{n} + \left(\frac{\mathbf{e}'\mathbf{D}\mathbf{e}}{n}\mathbf{I} - 2\mathbf{D}\right)\frac{\mathbf{e}}{n}.$$

The similarity between the structure of the two expressions is obvious. The term $\left(\frac{\mathbf{e}'\mathbf{D}_k\mathbf{e}}{n}\mathbf{I} - 2\mathbf{D}_k\right)\frac{\mathbf{e}}{n}$ represents the squared distances of the points that generate the distances of $\mathbf{D}_k$ from their centroid, G_k, say. When this term is used to approximate $\{f(0,x_{ik})\}$, the two expressions become identical. With linear biplots the equivalence is exact because then O and G coincide; it may often give a good approximation, but this possibility needs further examination.

In section 3, it was shown that the trajectories of non-linear biplots are concurrent at O. In section 4 it was pointed out that the trajectories for generalised biplots are not concurrent, but in 4.2 it was shown how they can be adjusted to concurrency for the purposes of interpolation. This adjustment is not valid for prediction, because it affects the nearness properties associated with the pseudo-samples. That the difference between (17) and (19) does not depend on ξ accounts for the fact that the nearness property holds for both. This result entails that for a given point P_i, the difference between the vectors joining P_i to the same marker ξ on the two trajectories is a constant vector, and in this sense, the two trajectories are parallel. This in turn means that the normals from P_i to the two trajectories cannot have the same marker and so must give different predictions, which is an apparent anomaly. The problem is resolved by noting that the extra dimensions given by (13) correct for the difference. Thus, these dimension plays an important part in determining the correct back-projections for prediction. It might be worth examining new forms of pseudo-sample in the hope of finding one that simplifies the properties of the additional dimensions. One possibility suggested by (17) and (19) is to take $\{f(\xi,x_{ik})\}$ as the vector of squared distances between a pseudo-sample and the sample-points. This certainly preserves the nearness property and it is more simple than the vectors (17) and (19) given by the two forms of pseudo-sample so far considered. The main problem with it is the lack of an expression for the associated pseudo-sample and there is no guarantee that such exists without recourse to complex numbers. The advantage of the pseudo-samples used with non-linear and generalised biplots, is that they have explicit forms that correspond to possible real samples, however unlikely it is that they would occur in practice. Assumption (i) of section 3 guarantees that the pseudo-sample has a real representation in $\mathfrak{R}_n$.

6 Conclusion

In the above, it has been assumed that the samples have an exact representation in at most n-1 dimensions and it has been shown that they can be represented relative to coordinate axes, possibly non-linear, in at most n^* dimensions. Of course, in the linear case both dimensions are reduced to p. Biplot axes have been defined as representations of these coordinate axes when the configuration of sample-points is approximated in $r<p<n^*$ dimensions. It has been shown that the form that these axes take in the approximation space depends critically on whether they are to be used for interpolating new samples into that space or are to be used for predicting the values of the original variables for existing or putative samples. Both for prediction and interpolation, the biplot axes may be treated very much as conventional coordinate axes so long as appropriate scales with associated markers are provided. Interpolation is in terms of forming vector-sums. Prediction is in terms of various ways of projecting points onto axes, which may be thought of as finding the nearest (which may be distant) point on an axis to a sample-point. These are seen as the main tools for interpreting biplots. The situation is much the same for categorical variables, where the biplot axes become sets of discrete CLPs. Interpolation remains a matter of vector-sums and prediction is now overtly in terms of nearness, leading to the consideration of neighbour-regions.

One of the reasons why there is so much complication, is that for prediction purposes two objectives are confounded. The primary objective for prediction is the calculation of $\mathfrak{N} \cap \mathfrak{L}$ for a convenient set of markers in $\mathfrak{M}$. Even in the linear case, this would result in the unacceptable clutter of sets of parallel lines in $\mathfrak{L}$ for every variable. Consequently a secondary objective is to characterise the intersection in some simple way. The answer is to appeal to orthogonal extension, which allows all the parallel lines for one variable to be replaced by a single biplot axis, in the knowledge that the full intersection space at any point on the axis can be reconstructed as the space normal to the given point in $\mathfrak{L}$. In section 3.3 it was shown that a similar simplification is available for non-linear biplots and that, for continuous variables, this carried over into generalised biplots. The intersection space for categorical variables normally requires at least two dimensions to convey even an approximation of the spatial properties of neighbour-regions. When, as is usual, $\mathfrak{L}$ is two-dimensional, this leaves no room for simplification, nor is simplification important, because normally there are only a few prediction-regions. In three-dimensional approximations only two-dimensional neighbour-regions (variables with three category levels) would offer scope for orthogonal extension but even then the outcome would be exceedingly complicated to use. Thus, for continuous variables the main thrust is on the simplification objective, while for categorical variables it is on displaying the intersection space without simplification. The prediction-regions in $\mathfrak{L}$ are neighbour regions for the CLPs in $\mathfrak{M}$ and, as pointed out in section 4.3, the back-projections of the CLPs onto $\mathfrak{L}$ generate only approximate neighbour-regions within $\mathfrak{L}$. What one would like is a set of prediction-CLPs in $\mathfrak{L}$ that generate neighbour-regions in $\mathfrak{L}$ that coincide with those described above. It is not known whether such exist, but if they do, they would provide the simplification required for exhibting all categorical variables simultaneously in a

manner that allowed prediction, in a similar way that the orthogonally projected CLPs permit interpolation within $\mathbf{L}$.

Those familiar with the biplot literature may be surprised to find no mention in this paper of correlation and little mention of principal axes. With linear biplots it is true that among all possible projections, the projection of the sample points onto β_k gives a set of values that have maximal correlation with the sample-values of the kth variable. This certainly gives a characterisation of β_k but, in my opinion, one that is not of great interest. After all, projections onto the original axis ξ_k give unit correlation, whatever the values of the kth variable. One may interpret the result as showing that β_k is the best representation of ξ_k in the approximation space $\mathbf{L}$ and this leads one again into interpreting β_k as a coordinate axis. Regarding principal axes, it has been shown that the main results are independent of the choice of $\mathbf{L}$. Nevertheless, it is natural to choose $\mathbf{L}$ to give a best representation of the samples and this often leads to eigenvalue problems that define $\mathbf{L}$ in terms of spanning eigenvectors. Then, in addition to the biplot axes, the representation of $\mathbf{Y}$ in $\mathbf{L}$ is also with respect to r orthogonal principal axes. Those who favour the reification of principal axes may continue to reify and they may find the biplot axes a useful guide to interpretation of principal axes or, indeed, they may find that the biplot axes are helpful in suggesting interesting sets of oblique axes.

Finally, the analysis presented above has been very much in terms of projections. Yet Gower and Harding (1988) and Gower (1991) have pointed out that the pseudo-sample idea for both non-linear and generalised biplots may easily be modified for any form of metric or non-metric scaling and this has been implemented by Underhill and by Heiser and Meulman, in so far unpublished work. Provided the new points are added by a method that is coherent with the criterion used for constructing the multidimensional scaling (see Gower, 1991), the interpolation methods discussed above should continue to be satisfactory but it is less clear what becomes of prediction, which is so intimately related to the concepts of back-projection and nearness. Even the notion of the two spaces $\mathbf{L}$ and $\mathfrak{M}$ raises problems, though perhaps progress may be made here by embedding in $\mathfrak{R}$ the multidimensional scaling approximation $\mathbf{L}$, using orthogonal Procrustes analysis. This proposal has the advantage that principal coordinate, and hence also principal component, approximations are optimally oriented to the configuration of samples in the full space, as is easily seen because the Procrustes residual is the same as the minimised component sum-of-squares of the residuals orthogonal to $\mathbf{L}$. Automatically, $\mathfrak{M}$ will be part of this embedding but as r varies so does $\mathfrak{M}$ and a problem has to be resolved in understanding how the different forms of $\mathfrak{M}$ are related. The results presented here may be viewed as a gauge (Gifi, 1990) for what to expect, or at least what needs further investigation, within this more general context.

References

Anderson T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley.

Eckart C. and Young G. (1936). The approximation of one matrix by another of lower rank *Psychometrika*, **1**, 211-8.

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

Gifi A. (1990). *Non-linear Multivariate Analysis*. New York: J. Wiley and Son.

Gower J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325-38.

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

Gower J.C. (1982). Euclidean distance analysis. *The Mathematical Scientist*, **7**, 1-14.

Gower J. C. (1991). Generalised biplots. Research Report RR-91-02. Leiden: Department of Data Theory.

Gower J. C. (1992). The construction of neighbour-regions in two dimensions for prediction with multi-level categorical variables.

Gower J. C. and Legendre P. (1986). Metric and Euclidean Properties of Dissimilarity Coefficients. *Journal of Classification*, **3**, 5-48.

Gower J. C. and Harding S. (1988). Non-linear biplots. *Biometrika*, **73**, 445-55.

Greenacre M.J. (1984). *Theory and Applications of Correspondence Analysis*. London: Academic Press

Torgerson W. S. (1955). *Theory and Methods of Scaling*. New York: John Wiley.

Appendix A

Derivation of Algebraic Formulae and Other Results for Back-Projection

Basic Formulae

Figure 2 gives a geometrical representation of an important special case of back-projection. In general we are given a point $\mathbf{x}$ in $\mathfrak{R}_p$ and two sub-spaces: $\mathfrak{L}$ of dimension r and $\mathfrak{M}$ of dimension s. An algebraic expression is required for the point $\mathbf{y} \in \mathfrak{N} \cap \mathfrak{L}$ that is closest to $\mathbf{x} \in \mathfrak{N}$, where $\mathfrak{N}$ is normal to $\mathfrak{M}$. We shall refer to $\mathbf{y}$ as the back-projection of $\mathbf{x}$ in $\mathfrak{L}$ through $\mathfrak{M}$; when $\mathbf{x} \in \mathfrak{M}$, $\mathbf{y}$ is simply termed the back-projection of $\mathbf{x}$ in $\mathfrak{L}$. Indeed, provided $\mathfrak{L}$ and $\mathfrak{M}$ are allowed to be disjoint spaces, the concept of back-projection suffices, because it can always be arranged that $\mathbf{x} \in \mathfrak{M}$. Thus it is assumed that $\mathfrak{L}$ contains the origin G, as before, while $\mathfrak{M}$ has an offset $\mathbf{q}$ at a point marked Q. Rather than choose Q arbitrarily, fix it as the projection of G onto $\mathfrak{M}$; thus $\mathbf{q} = \mathbf{xNN}'$. Figure 9 illustrates the geometry of this generalisation.

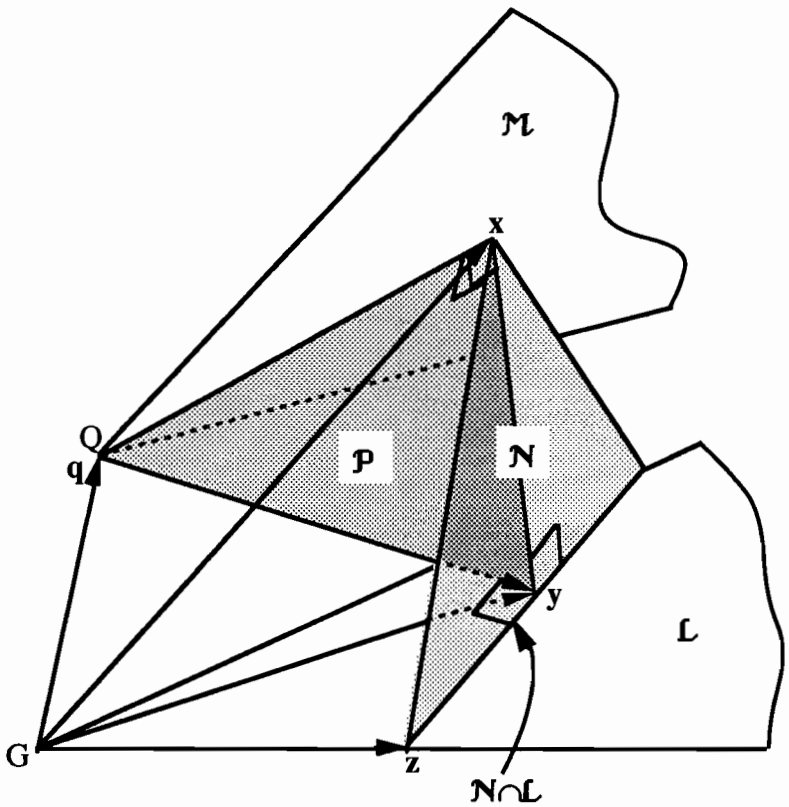

Figure 9. This is a generalisation of Figure 2, in which the spaces $\mathfrak{L}$ and $\mathfrak{M}$ are now allowed to be disjoint with an offset $\mathbf{q}$. $\mathfrak{N}$ is the space normal to $\mathfrak{M}$ at $\mathbf{x}$ and $\mathbf{y}$ is the back-projection of $\mathbf{x}$ in $\mathfrak{L}$ and hence lies in $\mathfrak{N} \cap \mathfrak{L}$. $\mathbf{z}$ is a general point in $\mathfrak{N} \cap \mathfrak{L}$. Several orthogonalities are indicated.

In this formulation, the origin G is an arbitrary point in $\mathfrak{L}$ so it follows that $\mathbf{q}$ is equally arbitrary in $\mathfrak{M}$. Thus, although $\mathbf{q} = \mathbf{0}$ certainly implies that $\mathfrak{M}$ and $\mathfrak{L}$ intersect, the converse is not true when $\mathbf{q} \neq \mathbf{0}$. Thus, to refer to $\mathbf{q}$ as an offset could be misleading. The difficulty can be avoided by choosing the origin to be at the point in $\mathfrak{L}$ that is nearest $\mathfrak{M}$, which would define $\mathbf{q}$ to be a true offset when $\mathfrak{L}$ and $\mathfrak{M}$ did not intersect and to be null when they do. This course has not been followed, because in our applications of the results given in this appendix, the origin G is more naturally defined as the mean of n sample-points in $\mathfrak{L}$, which is an arbitrary point in terms of the geometry described.

Writing $\mathfrak{K}$ for the space normal to $\mathfrak{L}$, that $\mathbf{y} \in \mathfrak{N} \cap \mathfrak{L}$ may be expressed as:

$$\text{(i) } \mathbf{yMM'} = \mathbf{xMM'} \quad \text{and} \quad \text{(ii) } \mathbf{yKK'} = \mathbf{0}, \tag{A1}$$

where (i) is valid for any $\mathbf{y} \in \mathfrak{N}$ and (ii) is valid for any $\mathbf{y} \in \mathfrak{L}$. Thus we require the minimum of $(\mathbf{y} - \mathbf{x})(\mathbf{y} - \mathbf{x})'$ with respect to $\mathbf{y}$ and subject to the constraints (A1). Introducing Lagrange multipliers in two p-dimensional row-vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ gives, after differentiation:

$$\mathbf{y} = \mathbf{x} + \boldsymbol{\lambda}\mathbf{KK'} + \boldsymbol{\mu}\mathbf{MM'}. \tag{A2}$$

Multiplying (A2) by $\mathbf{NN'}$ and using (i) of (A1) yields:

$$\mathbf{y} - \mathbf{x} = \boldsymbol{\lambda}\mathbf{KK'NN'} \tag{A3}$$

and hence:

$$\boldsymbol{\lambda}\mathbf{K} = (\mathbf{y} - \mathbf{x})\mathbf{K}(\mathbf{K'NN'K})^{-1},$$

which on using (ii) simplifies to :

$$\boldsymbol{\lambda}\mathbf{K} = -\mathbf{x}\mathbf{K}(\mathbf{K'NN'K})^{-1}, \tag{A4}$$

provided the inverse exists, which will usually be satisfied when $r \geq s$.
Similarly, eliminating $\boldsymbol{\lambda}$ from (A2) gives:

$$\boldsymbol{\mu}\mathbf{M} = \mathbf{xKK'M}(\mathbf{M'LL'M})^{-1}. \tag{A5}$$

Substituting (A4) and (A5) into (A2) yields after a considerable amount of algebraic manipulation:

$$\mathbf{y} = \mathbf{x}(\mathbf{I} - \mathbf{K}(\mathbf{K'NN'K})^{-1}\mathbf{K'NN'}), \tag{A6}$$

which is equation (4). An alternative expression is:

$$\mathbf{y} = \mathbf{x}(\mathbf{I} + \mathbf{KK'M}(\mathbf{M'LL'M})^{-1}\mathbf{M'})\mathbf{LL'}. \tag{A7}$$

This is the basic result (5) which, it can easily be checked, satisfies the conditions (A1) and can be rewritten in the form (A2). Both (A6) and (A7) are normally valid for $r \geq s$.

From $\mathbf{q} = \mathbf{xNN'}$ we have that $\mathbf{x} = \mathbf{xMM'} + \mathbf{q}$. Substituting for $\mathbf{x}$ in (A7) then gives:
$\mathbf{y} = \mathbf{xMM'LL'} + \mathbf{xM}(\mathbf{M'KK'M})(\mathbf{M'LL'M})^{-1}\mathbf{M'LL'} + \mathbf{qLL'} + \mathbf{qKK'M}(\mathbf{M'LL'M})^{-1}\mathbf{M'LL'}$
which, on substituting $\mathbf{I} - \mathbf{LL'}$ for $\mathbf{KK'}$ and recalling that $\mathbf{qM} = \mathbf{0}$, simplifies to:

$$\mathbf{y} = \mathbf{xM}(\mathbf{M'LL'M})^{-1}\mathbf{M'LL'} + \mathbf{qLL'}(\mathbf{I} - \mathbf{M}(\mathbf{M'LL'M})^{-1}\mathbf{M'LL'}) \tag{A8}$$

This alternative form to (A6) and (A7) is not the most convenient for practical use but it shows that in the very important special case when $\mathbf{q} = \mathbf{0}$, (A6) and (A7) simplify to:

$$\mathbf{y} = \mathbf{xM}(\mathbf{M'LL'M})^{-1}\mathbf{M'LL'}. \tag{A9}$$

Indeed, for (A9) to be valid, it suffices that $\mathbf{qL} = \mathbf{0}$, a condition that fails to be satisfied when $\mathfrak{M}$ is given by the space of category-level points (CLPs) of section 4.3.

Orthogonality Results

Let **x**, **y** and **z** be vectors as defined above and summarised in Figure 9. Thus $\mathbf{x} \in \mathcal{M}$, **y** given by (A7) is the back-projection of **x** in $\mathcal{L}$ and **z** is any other vector in $\mathcal{N} \cap \mathcal{L}$. It is now shown that the followoing pairs of vectors are orthogonal:

(a) $(\mathbf{x} - \mathbf{z}) \perp (\mathbf{x} - \mathbf{q})$ (b) $(\mathbf{x} - \mathbf{y}) \perp (\mathbf{x} - \mathbf{q})$ (c) $(\mathbf{y} - \mathbf{z}) \perp (\mathbf{x} - \mathbf{q})$. (A10)

In all these results **q** may be replaced by its equivalent **xNN'**. The proof is almost by definition. Because $\mathcal{N}$ is normal to $\mathcal{M}$ at **x** it follows that for all $\mathbf{z} \in \mathcal{N}$ we have that $(\mathbf{x} - \mathbf{z}) = (\mathbf{x} - \mathbf{z})\mathbf{NN}'$. Because $\mathcal{N}$ is orthogonal to $\mathcal{M}$ we have $\mathbf{N}'\mathbf{M} = \mathbf{0}$. Combining these results gives $(\mathbf{x} - \mathbf{z})\mathbf{MM}'\mathbf{x}' = (\mathbf{x} - \mathbf{z})\mathbf{NN}'\mathbf{MM}'\mathbf{x}' = 0$, thus proving (a) of (A10). The result (b) immediately follows by substituting **y** as a special case of the general vector **z**, and (c) follows as the difference between (a) and (b). The following results are of key importance for the success of biplot techniques:

(d) $(\mathbf{y} - \mathbf{z}) \perp (\mathbf{y} - \mathbf{q})$ and (e) $(\mathbf{y} - \mathbf{z}) \perp (\mathbf{y} - \mathbf{x})$. (A11)

To prove (d), substitute (A7) for **y** to give:

$$(\mathbf{y} - \mathbf{z})(\mathbf{y} - \mathbf{q})' = (\mathbf{xB} - \mathbf{z})(\mathbf{xB} - \mathbf{xNN}')'$$
$$\text{where } \mathbf{B} = (\mathbf{I} + \mathbf{KK}'\mathbf{M}(\mathbf{M}'\mathbf{LL}'\mathbf{M})^{-1}\mathbf{M}')\mathbf{LL}'.$$

Expanding gives:

$$(\mathbf{y} - \mathbf{z})(\mathbf{y} - \mathbf{q})' = \mathbf{xBx}' - \mathbf{zBx}' - \mathbf{xBNN}'\mathbf{x}' + \mathbf{zNN}'\mathbf{x}'$$

which, on using $\mathbf{zK} = 0$ (because $\mathbf{z} \in \mathcal{L}$) and $\mathbf{zMM}' = \mathbf{xMM}'$ (because $\mathbf{z} \in \mathcal{N}$), simplifies to

$$(\mathbf{y} - \mathbf{z})(\mathbf{y} - \mathbf{q})' = \mathbf{xBMM}'\mathbf{x}' - \mathbf{zx}' + (\mathbf{z} - \mathbf{xMM}')\mathbf{x}'$$

which on substituting for **B** gives

$$(\mathbf{y} - \mathbf{z})(\mathbf{y} - \mathbf{q})' = \mathbf{xLL}'\mathbf{MM}'\mathbf{x}' + \mathbf{xKK}'\mathbf{MM}'\mathbf{x}' - \mathbf{xMM}'\mathbf{x}' = 0,$$

establishing the result. The difference between (d) and (c) gives (e).

A geometrical derivation and discussion of these results is given in section 2.4 and illustrated in Figure 4. Result (d) is important because it does not involve **x**. When the two spaces share the origin, so that $\mathbf{q} = \mathbf{0}$, this shows that given the back-projection **y**, the whole of the intersection space $\mathcal{N} \cap \mathcal{L}$ may be constructed by an orthogonal extension of **y** into $\mathcal{L}$, that is by using only that part of the normal space to **y** that is contained in $\mathcal{L}$; this requires no knowledge of **x**. When $\mathbf{q} \neq 0$ the extension into $\mathcal{L}$ from **y** is oblique, which is an inconvenience. This inconvenience can be avoided by finding the pointy $\mathbf{y}^*$ in $\mathcal{N} \cap \mathcal{L}$ that is nearest G. To find $\mathbf{y}^*$ note that the projection matrix for $\mathcal{N} \cap \mathcal{L}$ is $\mathbf{I} - \mathbf{RR}'$, say, given by:

$$\mathbf{RR}' = \mathbf{KK}' + \mathbf{LL}'\mathbf{M}(\mathbf{M}'\mathbf{LL}'\mathbf{M})^{-1}\mathbf{M}'\mathbf{LL}'. \quad \text{(A12)}$$

The vectors **R** are normal to $\mathcal{N} \cap \mathcal{L}$ so $\mathbf{y}^*$ is given by $\mathbf{yRR}'$, the projection of **y** onto this normal space. Substituting (A6) and (A12) for **y** and **RR'**, respectively, gives after some algebraic manipulation:

$$\mathbf{y}^* = \mathbf{xM}(\mathbf{M}'\mathbf{LL}'\mathbf{M})^{-1}\mathbf{M}'\mathbf{LL}' \quad \text{(A13)}$$

which is identical to (A9). The explanation is that $\mathbf{y}$ is the point in $\mathcal{N} \cap \mathcal{L}$ that is nearest $\mathbf{x}$ while $\mathbf{y}^*$ is the point in $\mathcal{N} \cap \mathcal{L}$ that is nearest G; when $\mathbf{q} = \mathbf{0}$, and as was shown above $\mathbf{qL} = \mathbf{0}$, $\mathbf{y}$ and $\mathbf{y}^*$ coincide. Both (A7) and (A9) have their uses.

Back-projection has been defined as determining a point $\mathbf{y}$ when given a point $\mathbf{x}$. Clearly agglomerations of points $\mathbf{x}$ into lines, higher-dimensional spaces or neighbour-regions, determine higher-dimensional spaces of back-projections.

Deficient Rank Situations

When $s > r$ the matrices in (A6) and (A7) that require inversion are of deficient rank and the formulae are then certainly invalid. The problem is that the dimensionality of $\mathcal{N} \cap \mathcal{L}$ cannot exceed $r - s$, so only a part $\mathcal{M}^*$ of the higher-dimensional space $\mathcal{M}$ can back-project into $\mathcal{L}$. Indeed, because if $\mathbf{y}$ is the back-projection in $\mathcal{L}$ of any $\mathbf{x} \in \mathcal{M}$ then $\mathbf{x}$ is the projection of $\mathbf{y}$ onto $\mathcal{M}$, it follows that $\mathcal{M}^*$ is the orthogonal projection of $\mathcal{L}$ onto $\mathcal{M}$. The remainder of the intersection $\mathcal{N} \cap \mathcal{L}$, if required, may be constructed by orthogonal extension as described immediately above. This situation is briefly explored in section 4.3 and addressed more fully in Gower (1992).

Even when $s \leq r$, the above formulae become degenerate whenever $\text{rank}(\mathbf{L}'\mathbf{M}) < s$ which is when a subspace of $\mathcal{M}$ is orthogonal to $\mathcal{L}$. Again $\mathcal{M}^*$ is the space common to $\mathcal{L}$ and $\mathcal{M}$ and we may proceed as in the previous paragraph.

Appendix B

Simultaneous Representation of p -1 Axes by Back-Projection onto Two Dimensions

In this appendix rather than expressing results relative to $\mathfrak{R}_p$, we shall represent all coordinates with respect to axes in the r -dimensional space of $\mathfrak{L}$. Thus (8) now becomes;

$$\mathbf{y} = \mathbf{x}\frac{\mathbf{mm'L}}{\mathbf{m'LL'm}}$$

and when $\mathbf{m} = \mathbf{e}_k'$ the above simplifies to:

$$\mathbf{z}_k = \frac{\{\mathbf{L}\}_k}{(\mathbf{LL'})_{kk}} \tag{B1}$$

where $\{\mathbf{L}\}_k$ is the kth row of $\mathbf{L}$ and $\mathbf{z}_k$ is the back-projection of the unit marker of ξ_k.

With r =2, let $\mathbf{L} = \begin{pmatrix} u_1 & v_1 \\ u_2 & v_2 \\ u_3 & v_3 \\ .. & .. \\ u_p & v_p \end{pmatrix}$, then from (B1) $\mathbf{z}_k = \frac{(u_k \; v_k)}{(u_k^2 + v_k^2)}$. (B2)

$\mathfrak{L}$ is to be chosen so that $\mathbf{z}_1 = \mathbf{z}_2 = \mathbf{z}_3 = ... = \mathbf{z}_{p-1}$ which implies that the first p -1 rows of $\mathbf{L}$ are constant, i.e. $\mathbf{L}$ has the form::

$$\mathbf{L} = \begin{pmatrix} u & v \\ u & v \\ u & v \\ .. & .. \\ u & v \\ u_p & v_p \end{pmatrix}. \tag{B3}$$

Then the conditions for orthonormality give:

$$u_p^2 = 1 - (p-1)u^2, \qquad v_p^2 = 1 - (p-1)v^2, \qquad u_p v_p + (p-1)uv = 0, \tag{B4}$$

which for consistency requires that $u^2 + v^2 = \frac{1}{p-1}$, and so simplifying (B4) to:

$$u_p^2 = 1 - (p-1)u^2, \qquad v_p^2 = (p-1)u^2 \tag{B5}$$

where opposite signed square roots must be taken when calculating u_p and v_p .

For arbitrary choices of u, the settings (B5) ensure that the markers $(\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_{p-1})$ all back-project to the same point $\mathbf{z}$= $(p - 1)(u, v)$; also, $\mathbf{e}_p$ back-projects to (u_p, v_p). This does not mean that the same values are predicted for all p-1 axes, because the unit marker on each axis may be selected independently to represent any convenient value in the actual scale of measurement. However, all sets of predictions will be proportional.

Possible special cases are

$$\text{(i) } u_p = u = \frac{1}{\sqrt{p}}, \; v_p = -\sqrt{\frac{p-1}{p}}, \; v = \frac{1}{\sqrt{p(p-1)}}$$

and $$\text{(ii) } u_p = v = 0, \; u = \frac{1}{\sqrt{p-1}}, \; v_p = 1.$$

These results are given mainly to illustrate the methodology but (i) might have some interest as a method for isolating size effects. Rather than make p -1 axes have a common back-projection in two dimensions, it might be more useful to choose $\mathfrak{L}$ so that only some subset of axes are treated in this way, the remaining being chosen to optimise approximation in $\mathfrak{L}$.