

THE CONSTRUCTION OF NEIGHBOUR-REGIONS IN TWO
DIMENSIONS FOR PREDICTION WITH MULTI-LEVEL
CATEGORICAL VARIABLES.

John C. Gower

THE CONSTRUCTION OF NEIGHBOUR-REGIONS IN TWO DIMENSIONS FOR PREDICTION
WITH MULTI-LEVEL CATEGORICAL VARIABLES

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

Summary

In the multidimensional scaling of samples described by categorical variables, each l -level categorical variable is represented by a set of l points forming the vertices of a simplex. Any sample that has level i of a categorical variable will be nearer the corresponding vertex C_i than to any other vertex of the simplex, so defining convex neighbour-regions for each level. In r -dimensional approximations, predictions of levels are obtained by examining the intersections of the neighbour-regions with the r -dimensional space. The case $r = 2$ is of special practical importance and is the main concern of this paper; the methodology easily generalises. Examples are given of the forms taken by the neighbour-regions in planar sections of the $(l-1)$ -dimensional simplex and an algorithm is proposed for their construction.

1. Introduction

1.1 Geometry

Gower (1992) has discussed the geometry of linear biplots (Gabriel, 1971) of non-linear biplots (Gower and Harding, 1988) and of generalised biplots (Gower, 1991). Linear and non-linear biplots refer only to quantitative variables but generalised biplots also admit categorical variables. The k th categorical variable is represented by a set of l_k category level points (CLP) which form an s -dimensional simplex that is contained in a subspace \mathcal{M} of \mathcal{R}_n . Usually $s = l_k - 1$ but in special cases, such as with ordered categorical variables, we may have $s < l_k - 1$. Another subspace of \mathcal{R}_n is an r -dimensional space \mathcal{L} which contains the coordinates \mathbf{Y} of n points representing samples. We shall assume that $\mathbf{e}'\mathbf{Y} = 0$, so that the centroid G of the sample-points is at the origin; normally \mathbf{Y} will have been obtained from some form of multidimensional scaling. Under special conditions, G is also a point in \mathcal{M} so that \mathcal{L} and \mathcal{M} are then not disjoint, even though both may be small sub-spaces of \mathcal{R}_n . However, usually \mathcal{L} and \mathcal{M} are disjoint, although the offset between the two spaces is likely to be small - see Gower (1992) for more information on the roles of the spaces \mathcal{L} and \mathcal{M} .

Denote the q th vertex of the simplex by C_q ($q = 1, 2, \dots, l_k$). Gower (1991) gives simple formulae for calculating the coordinates of these vertices under different assumptions for defining distance with categorical variables; in the most simple case, these coordinates may be taken to be unit points on l_k orthogonal axes. The whole of \mathcal{R}_n may be partitioned into l_k n -dimensional regions, the q th of which contains all the points which are nearer C_q than any other vertex. These regions will be termed neighbour-regions and the q th neighbour-region for the k th categorical variable will be denoted by Q_q (for $q = 1, 2, \dots, l_k$). Let F_{qt} be the $(n-1)$ -dimensional flat that separates all points nearest C_q from all points nearest C_t . F_{qt} defines two trivial convex regions and Q_q is the intersection of all F_{qt} ($t \neq q$) which, being the intersection of convex regions, is itself convex. It is easy to see that given the part \mathcal{M}_q , of Q_q that is within \mathcal{M} , then the whole of Q_q is found by extending orthogonally into the remaining $n-s$ dimensions of \mathcal{R}_n (i.e. every $\mathbf{x} \in \mathcal{M}_q$ extends into the space normal to \mathcal{M}_q at \mathbf{x}). When $s = l_k - 1$, the neighbour-regions \mathcal{M}_q are convex cones, all with vertex C at the circumcentre of the simplex. \mathcal{R}_n may be partitioned into neighbour-regions Q_q , ($q = 1, 2, \dots, l_k$), in as many ways as there are categorical variables but attention will be focussed on the k th variable. Hence, in the following, the suffix k will be dropped, except when it is wished to emphasize reference to the k th variable. The neighbour-region Q_q intersects \mathcal{L} and the q th category-level will be predicted for all sample points \mathbf{Y} that lie in the intersection, which will therefore be termed a prediction-region. The prediction-regions must themselves be convex, because they are intersections of a linear sub-space with the

convex regions Q_q ($q = 1, 2, \dots, l$). This paper is concerned with the construction of the prediction-regions within \mathcal{L} .

A simple operational method for constructing the prediction-regions is as follows. Imagine \mathcal{L} to be represented by a set of pixels. Associate a colour, not black, with each category-level. Then calculate the distance of pixel P from C_q (for $q = 1, 2, \dots, l$); generally, one of these distances will be the shortest, say that for level h . If the shortest distance is not unique, colour P black; otherwise, colour P with colour h . Do this for all pixels. In this way, the prediction-regions of \mathcal{L} will be shown as coloured convex regions with linear boundaries. Repeat for each categorical variable. This may suffice for many practical applications but the following explores the underlying geometry more deeply and offers an algorithm which should be more efficient, especially with many category-levels.

The orthogonal projection, \mathcal{M}^* , of \mathcal{L} onto \mathcal{M} gives that part, often the whole, of \mathcal{M} which may orthogonally extend into \mathcal{L} . Thus, the main problem in constructing the neighbour-regions Q_k is to find the intersections of \mathcal{M}^* with the neighbour-regions \mathcal{M}_k within \mathcal{M} . A secondary problem, the representation of these intersections in \mathcal{L} , depends on the relationships between the dimensions r , s and s^* , the dimensionality of \mathcal{M}^* . When $s^* < r$, an $(r - s^*)$ -dimensional subspace of \mathcal{L} projects into \mathcal{N}^* , a subspace of \mathcal{N} ($= \mathcal{R}_n - \mathcal{M}$), the space normal to \mathcal{M} . Then the representation of \mathcal{M}^* in \mathcal{L} has to be augmented by orthogonal extension into the whole of \mathcal{L} . This paper is mostly concerned with $r = 2$, the case of greatest practical importance, but also briefly examines $r = 1$. These two cases cover most of the variant situations and much of what follows generalises.

1.2 Algebra

The process of orthogonal extension requires the notion of what Gower (1992) termed back-projection. Back-projection of a point x in \mathcal{M} onto \mathcal{L} is defined to be the point $y \in \mathcal{N} \cap \mathcal{L}$ that is closest to x , where \mathcal{N} is normal to \mathcal{M} at x . Writing L and M for matrices whose columns give orthonormal bases for \mathcal{L} and \mathcal{M} , respectively, and K and N as their orthogonal complements, these conditions may be expressed as:

$$(i) \quad yMM' = xMM'$$

and $(ii) \quad yKK' = 0.$

Gower (1992) shows that provided $r \geq s$, the minimum of $(y - x)(y - x)'$ with respect to y subject to the constraints (i) and (ii) is given by:

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

An alternative expression is:

$$y = x(I - K(K'NN'K)^{-1}K'NN'). \quad (2)$$

When \mathcal{M} contains the origin, G , then (2) simplifies to:

$$y = xM(M'LL'M)^{-1}M'LL'. \quad (3)$$

In the following, the back-projection will be required of a point in \mathfrak{M}^* that is obtained by projecting \mathfrak{L} onto \mathfrak{M} . Thus if $\mathbf{x} \approx \mathfrak{M}^*$ we have that $\mathbf{x} = \mathbf{vL'MM}' + \mathbf{q}$ for some vector \mathbf{v} . Substitution into (1) gives after a little manipulation:

$$\mathbf{y} = \mathbf{vL'M(M'LL'M)^{-1}M'LL}'. \quad (4)$$

2. Examples

The geometry described in section 1.1 is difficult to visualise in its full generality. Thus in this section clarification is sought by studying a few simple special cases, which, as will be seen, point the way to a general solution to constructing prediction-regions in \mathfrak{L} .

2.1 Three category-levels

Suppose the k th variable is Colour, with levels Blue or Green or Red, (say), in which case $l_k = 3$ and $s = 2$. Thus the CLPs form a triangle, the vertices C_1, C_2 and C_3 of which may be alternatively labelled as Blue, Green, and Red. These three CLPs determine the two-dimensional space \mathfrak{M} . Within \mathfrak{M} the neighbour-regions $\mathfrak{M}_1, \mathfrak{M}_2$ and \mathfrak{M}_3 are bounded by the perpendicular bisectors of the sides of the triangle and these are concurrent at the circumcentre (henceforth referred to as a c-centre) as is shown in Figure 1. Extending these boundaries by back-projection into the space \mathfrak{N} normal to \mathfrak{M} gives the partition of \mathfrak{R}_n into the neighbour-regions $\mathfrak{Q}_1, \mathfrak{Q}_2$ and \mathfrak{Q}_3 that contain all the samples with the respective category-levels. The intersections of these neighbour-regions with \mathfrak{L} give the prediction-regions in \mathfrak{L} that predict the corresponding category-levels for the contained sample-points.

When $r \geq s$, equations (1), (2) and (3) allow the back-projections into \mathfrak{L} to be calculated for the CLPs. When $r = s = s^*$, as in Figure 1, it is necessary only to back-project $l_k + 1$ points, the l_k CLPs and their c-centre, which it is convenient to label with the name of the categorical variable itself. The prediction-regions are then easily constructed, because back-projection transforms mid-points into mid-points, so the boundaries of the prediction-regions can be constructed as shown in Figure 1. Samples which fall in the back-projected prediction-regions of \mathfrak{L} are predicted to have the corresponding labelled category-levels. Prediction-regions are not neighbour-regions for the back-projected CLPs and seem not necessarily to be neighbour-regions for any set of points in \mathfrak{L} .

It would be impracticable to exhibit the prediction-regions for all categorical variables simultaneously but the prediction-CLPs obtained as the back-projections of the CLPs in \mathfrak{M} might suffice for doing most predictions by eye. Then all that would be seen in a plot of \mathfrak{L} are the points for Colour (the back-projected c-centre), Blue, Red and Green, and similarly for other categorical variables for which $r = s$.

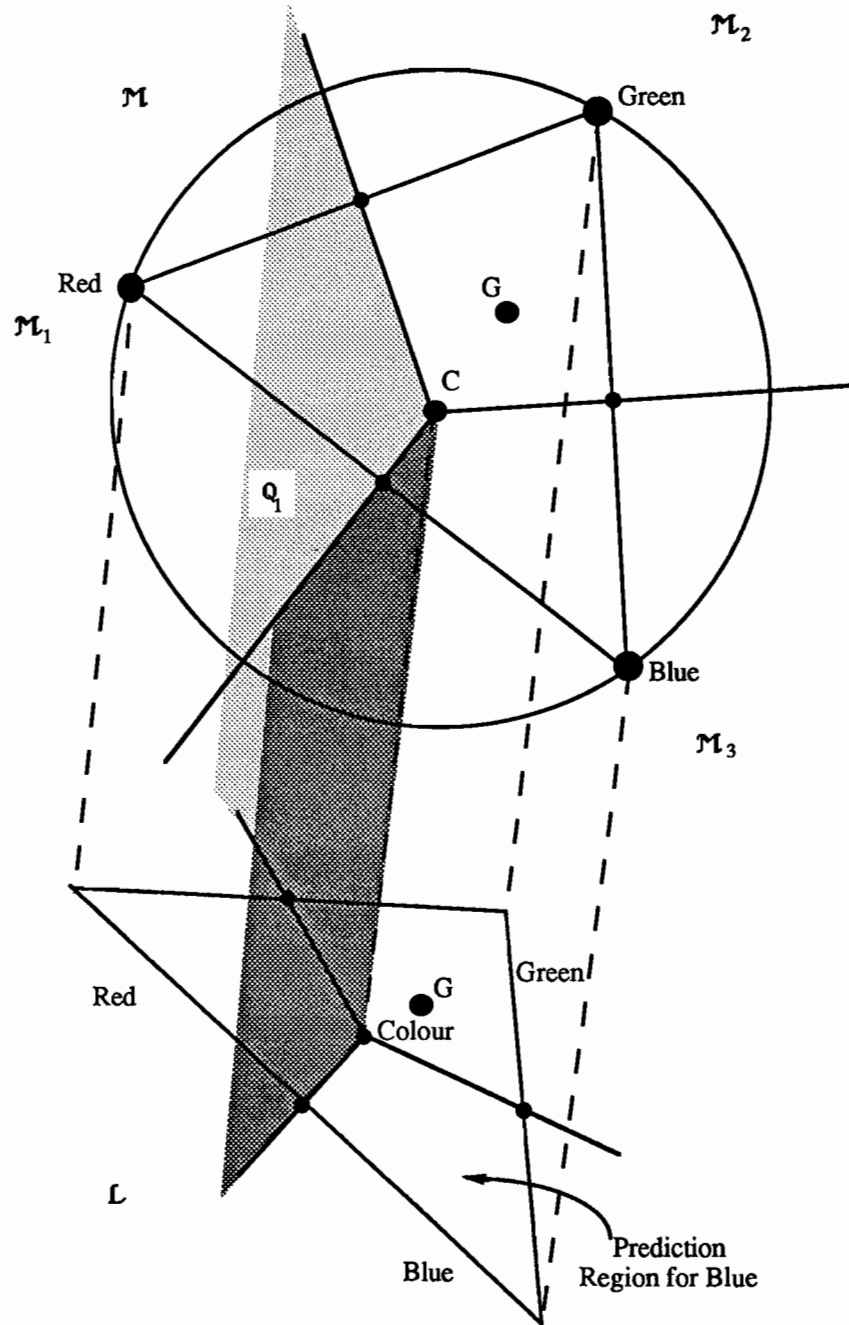


Figure 1. The case $r = s = s^* = 2$. For convenience \mathcal{L} and \mathcal{M} are shown as distinctly disjoint spaces although they are usually close and the centroid G may be common to both. The points for Red, Green and Blue in \mathcal{L} are the back-projections of the corresponding CLPs in \mathcal{M} , as are the mid-points which, together with the c-centre, determine the neighbour-regions in \mathcal{M} and the prediction-regions in \mathcal{L} . The partition of \mathcal{M} into neighbour-regions \mathcal{M}_1 , \mathcal{M}_2 and \mathcal{M}_3 is indicated, as is the extension into Q_1 to give the full neighbour-region for Red; the prediction region for Blue in \mathcal{L} is labelled. The dashed lines are orthogonal to \mathcal{M} .

When $r > s = s^*$, the only difficulty is that the back-projections define only an s -dimensional sub-space of \mathcal{L} . This is essentially the case of classical linear biplots, for

which $s = 1$ (a linear coordinate axis) and, usually, $r = 2$. The geometry discussed by Gower (1992) for linear biplots requires only minor modification to handle categorical variables. With categories, linearity corresponds to an ordered categorical variable. The back-projections of the CLPs will give l collinear (in general s -dimensional) points in \mathcal{L} and the prediction-regions can be completed by constructing the normals to the s -space within \mathcal{L} at the mid-points between the back-projected CLPs to give the back-projection of \mathcal{N}^* within \mathcal{L} (i.e. the part of the orthogonal extension of the boundaries of the \mathcal{M}_q that intersect \mathcal{L}). The geometry for the case $l = 4$, $s = s^* = 1$ and $r = 2$ is illustrated in Figure 2, where, for convenience, the mid-points, rather than the CLPs, have been back-projected immediately.

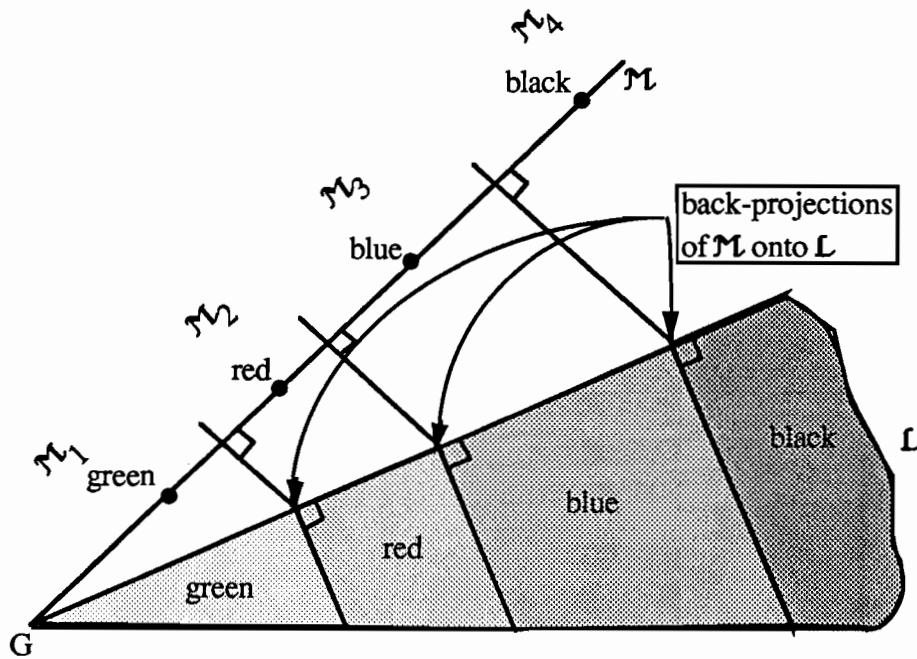


Figure 2. The prediction-regions for an ordered categorical variable with four levels (green, red, blue, black) in \mathcal{M} of $s = 1$ dimension represented in \mathcal{L} of $r = 2$ dimensions.

The case $s = 1$ is especially simple but when $1 < s < l_k - 1$, the problem of constructing the neighbour-regions of \mathcal{M} is itself non-trivial. The tessellation algorithm of (Bowyer, 1981 and Sibson, 1980) gives a solution for $s = 2$, $r \geq 2$, which then requires only a back-projection of the tessellation onto \mathcal{L} . The work of Devijver and Diekesei (1985) and of Watson (1981) promises extensions to higher values of s but the problem of constructing the intersection of the tessellation in \mathcal{M} with \mathcal{M}^* is more difficult than in the unrestricted case ($s = l_k - 1$) discussed in the following, and remains to be solved.

When $r < s$, the matrices in (1) and (2) that require inversion are of deficient rank and the formulae are invalid. Normally $r = 2$, so this difficulty arises whenever $l_k > 3$, which is a

common occurrence. First consider the case $s = 2$ and $r = s^* = 1$ illustrated in Figure 3. In this figure, the CLPs are the same as in Figure 1, so the neighbour-regions formed from back-projection of \mathcal{M} are identical. For simplicity the triangle is not shown but the mid-points of its sides on the lines separating the neighbour-regions are retained.

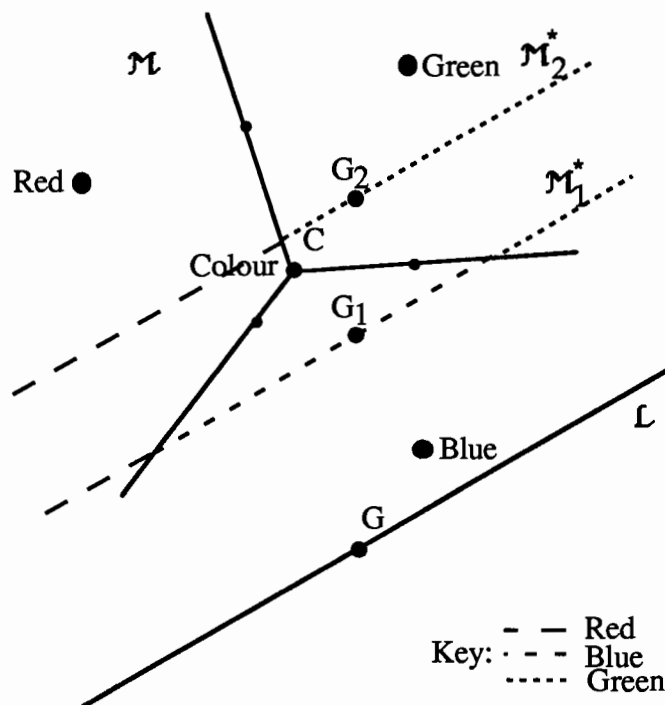


Figure 3. The case $s = 2$ and $r = s^* = 1$. The category-level points are as in Figure 1 but \mathcal{L} is here one-dimensional. Two possibilities are shown: (i) \mathcal{M}_1^* containing all three neighbour-regions and (ii) \mathcal{M}_2^* where neighbour-regions exist only for red and green. The back-projections onto \mathcal{L} , having the same structures, are not exhibited.

Now, however, most points in \mathcal{M} do not back-project into \mathcal{L} . Indeed, by definition, only points in \mathcal{M}^* will back-project into \mathcal{L} . In Figure 3 two possibilities, \mathcal{M}_1^* and \mathcal{M}_2^* , are shown, depending on the position, G_1 or G_2 , of the projection of the centroid G within the triangle with the vertices Red, Green and Blue. When \mathcal{M}_1^* passes through G_1 , there are neighbour-regions for all three category-levels. When \mathcal{M}_2^* passes through G_2 , only the neighbour-regions for the levels Red and Green intersect with \mathcal{M}^* . Normally none of the CLPs will now back-project into \mathcal{L} although, of course, there exist points in \mathcal{L} that are nearest to the category-level points; however, the projection onto \mathcal{M}_1^* of the CLP for Red falls into the Blue region showing that the prediction CLPs may fall into inconsistently labelled regions. It seems that for prediction, unlike for interpolation (see Gower, 1992), the projections of the CLPs onto \mathcal{L} may be misleading. This example shows that when $r < s$ there are special considerations. When $r = 2$, this situation will have to be faced whenever there are categorical variables with four, or more, levels.

2.2 Four Category-levels, Three Dimensions.

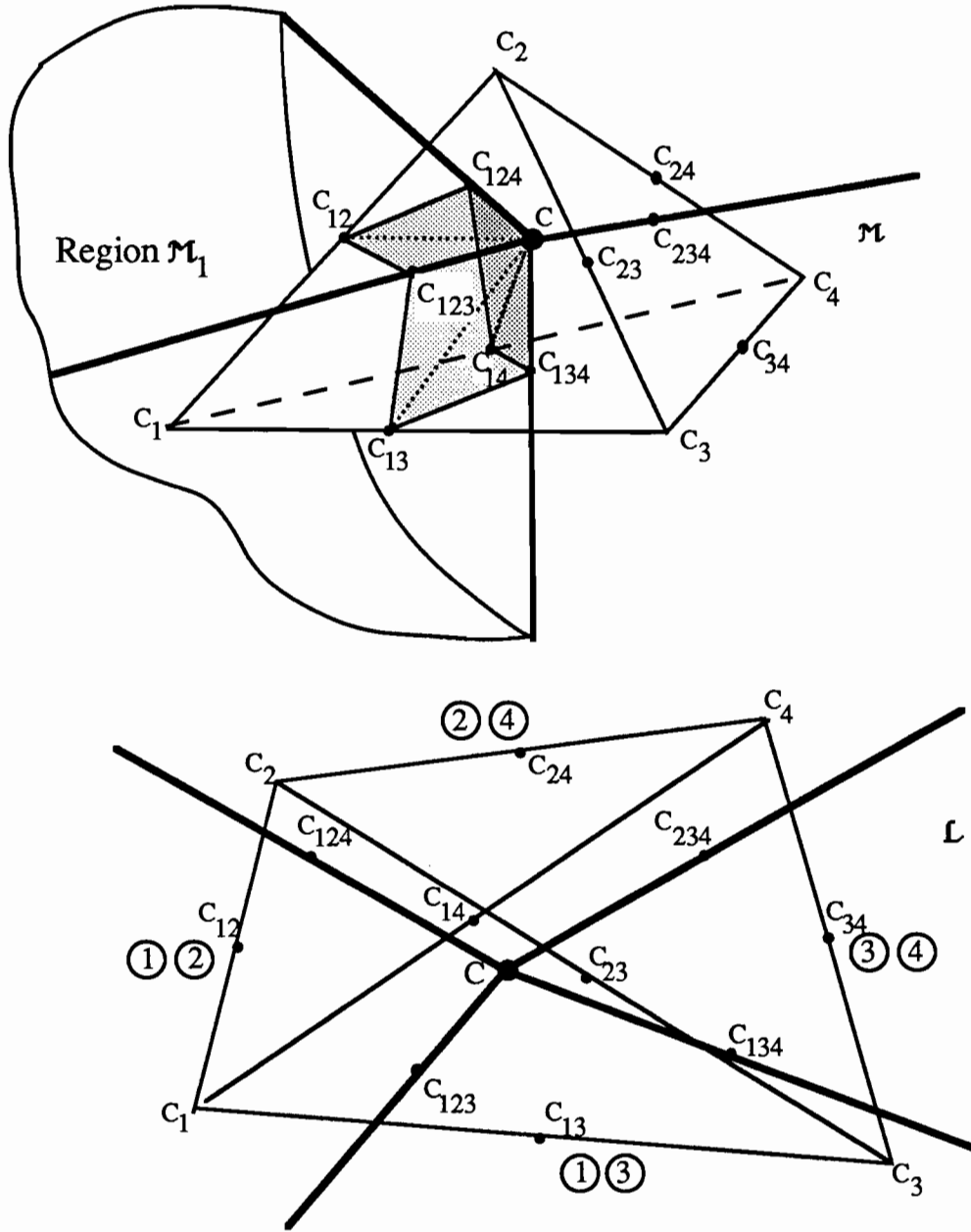


Figure 4. In \mathcal{M} the four category levels labelled C_1, C_2, C_3 and C_4 form a tetrahedron, c-centre C . The wedge-shaped neighbour-region \mathcal{M}_1 is shown for C_1 . The two-dimensional approximation in \mathcal{L} shows the orthogonal projections of the neighbour-regions. The edges joining C to the c-centres of the the four faces are shown bold.

With four levels, \mathcal{M} is three dimensional, with CLPs denoted by the vertices C_1, C_2, C_3 and C_4 of a tetrahedron, and the neighbour regions in \mathcal{M} are wedged-shaped, with plane boundaries whose edges are formed by the joins of the c-centre C with the c-centres of the triangles forming the four faces of the tetrahedron. For example, denoting the mid-point of C_iC_j by C_{ij} and the c-centre of $C_iC_jC_k$ by C_{ijk} then the region \mathcal{M}_1 has boundaries CC_{123} ,

CC_{124} and CC_{134} where, for example, the plane $CC_{123}C_{124}$ is normal to the edge C_1C_2 at C_{12} . First, to illustrate their deficiencies, we shall focus on the ordinary orthogonal projection of the tetrahedron onto \mathcal{L} of two dimensions and see what becomes of the neighbour-regions; this is shown in Figure 4.

These regions, unlike the back-projections to be discussed shortly, overlap as is indicated in the lower part of the figure, so that, as well as samples falling into wrongly labelled regions, every region is of uncertain prediction. The overlapping regions are shown more clearly in Figure 5 which is the same as Figure 4 but with the constructional lines removed. In Figure 5, the bold lines are the projections of lines joining C to the six mid-points.

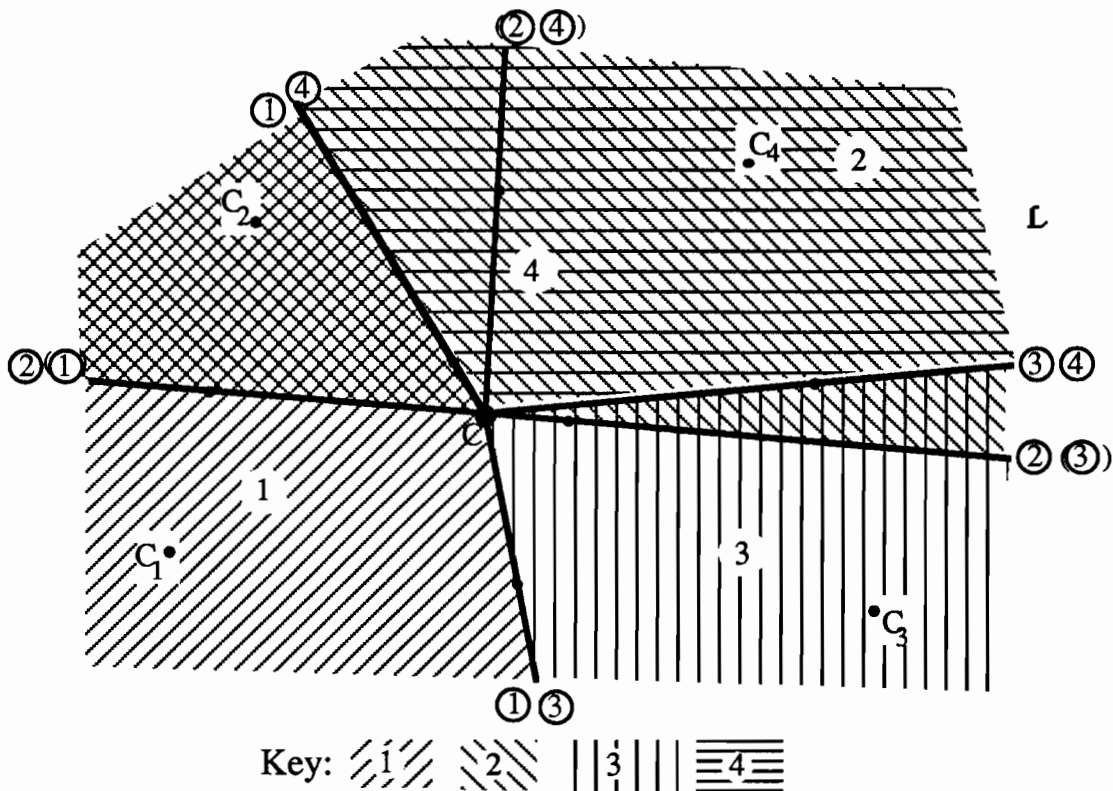


Figure 5. Figure 4 with the constructional details omitted to clarify the degree of overlap of the regions assigning to the four category-levels. Regions shown are projections not back-projections. Bracketed region labels indicate an overlapping boundary.

To have overlapping regions may seem unsatisfactory but in fact this situation is better than with numerical information, where each point of \mathcal{L} corresponds to all values that project from \mathcal{R}_n into that point. With categorical information, each region of \mathcal{L} is associated with only a subset of possible category-levels, as in Figure 5. However, as the number of levels increases, the overlapping regions become increasingly difficult to describe. Proper prediction requires the back-projection of \mathcal{M}^* onto \mathcal{L} . Because every point of \mathcal{R}_n belongs to a unique neighbour-region, this must give non-overlapping prediction-regions. In terms of Figure 4, we require the intersection of the neighbour-regions of the tetrahedron with \mathcal{M}^* . Just as with Figure 3, the resulting configuration depends on the position and the

orientation of the intersecting plane. There are three possible topologies depending on whether \mathcal{M}^* intersects with one, two, or three of the lines joining C to the four c-centres C_{ijk} ; four intersections are impossible, just as are three intersections in Figure 3. Intersections are at points on lines which separate the regions labelled \mathcal{M}_i , \mathcal{M}_j and \mathcal{M}_k , so will be at the meeting-point of three edges.

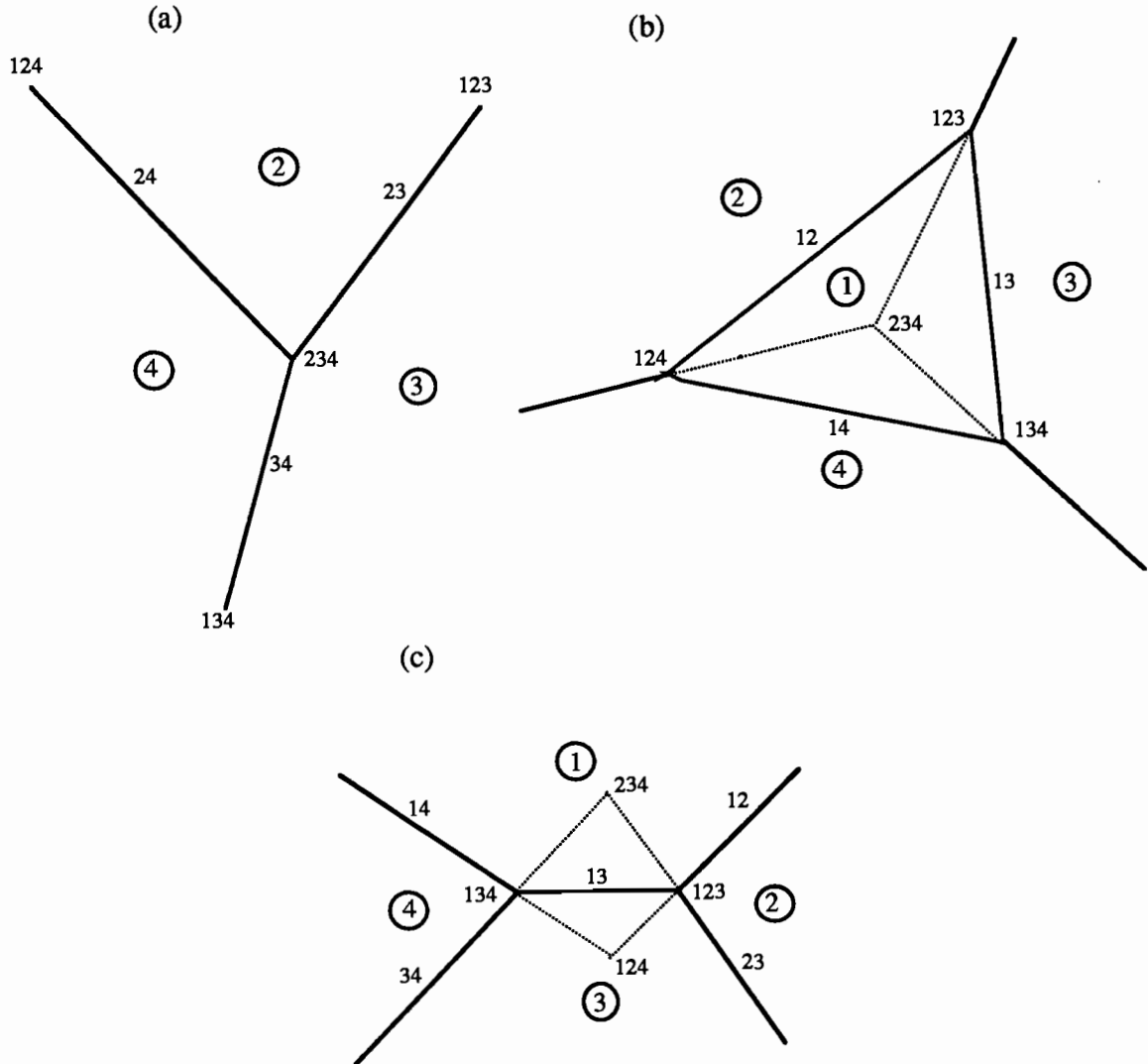


Figure 6. In (a) is shown the intersection with the neighbour-region \mathcal{M}_1 of a tetrahedron, formed by a slice through the region that excludes C_1 . In (b), the slice is through the space on the same side of C as is C_1 and cutting all three faces of \mathcal{M}_1 . In (c) the slice cuts two of the three edges of \mathcal{M}_1 .

Figure 6(a) shows a slice with one intersection, being a slice through CC_{234} on the opposite side of C as is \mathcal{M}_1 , while Figure 6(c) shows two intersections, arising from a slice on the other side of C, intersecting with all four regions. Figure 6(b) shows three intersections across the full range of \mathcal{M}_1 . In these figures an intersection with a line joining C to a c-centre of a face is denoted by a triplet ijk and to a mid-point by a doublet ij . When an intersection with ijk occurs by extending a line or plane beyond a limiting point, such as C, then dotted lines are used, as with 234 in figures 6(b) and 6(c); such a point is termed a virtual point. To understand virtual points better, consider the dotted line that joins 234 to

123 in Figure 6(b). This line separates regions \mathcal{M}_2 and \mathcal{M}_3 , so that any point on it is equally close to C_2 and C_3 but their nearest vertex is at C_1 . The line becomes undotted when it leaves \mathcal{M}_1 and the nearest vertex is then given by C_2 or C_3 .

Note that the neighbour-regions in one-dimensional approximations may be obtained by inspecting the intersection of a line with the different cases shown in Figure 6. In case (a) the situation is as in Figure 3, giving regions for either two or three levels; in cases (b) and (c), regions may be obtained that separate two, three or four levels.

2.3 Generalisations to Several Category-levels and Higher Dimensions

Next consider a two-dimensional section of a four-dimensional simplex, with vertices C_i ($i = 1,2,\dots,5$). Figure 7(a) is the analogue of Figure 6(b), being a three-dimensional intersection of the four-dimensional space, formed by slicing fully through \mathcal{M}_1 . In this figure, the three-dimensional nature of the regions cannot be labelled convincingly in a two-dimensional drawing. The device has been used of attaching a label to each arm that is to be associated with the region delimited by that arm, the two arms on either side (as seen in two-dimensions) and the face of the tetrahedron bounding \mathcal{M}_1 that is common to all three of these arms. Figure 7(b) reduces the dimensionality to two by slicing Figure 7(a) in its lower part towards the point marked 1245 but through \mathcal{M}_1 and avoiding \mathcal{M}_5 . It follows that all triplets involving the suffix 5 are missing; this is shown in 7(b) by placing the points 235, 245 and 345 at the extremities of the arms, indicating that they are at infinity on the corresponding lines and that \mathcal{M}_5 may also be regarded as if it were at infinity. Figure 7(c) includes four virtual points (134, 135, 145, 345) that coincide; these represent four lines which intersect somewhere on 1345, so this common-point is so identified and indicates that 7(c) is derived from a slice of the original four-dimensional simplex through \mathcal{M}_2 , on the opposite side of C as is CC_{1345} . Figures 7(d),7(e) and 7(f) show other possibilities.

The general pattern common to the variants shown in Figure 7 is now clear. All lines contain three triplets that share two suffices i,j (say). These lines are the intersections of \mathcal{M}^* , of dimension two, with the flat \mathcal{F}_{ij} of dimension $l-2$ that is normal to C_iC_j at its mid-point C_{ij} . (Recall that \mathcal{M}^* and \mathcal{M} are not disjoint spaces.) \mathcal{F}_{ij} contains all c-centres involving i and j , and, in particular, all c-centres of the form C_{ijk} . Indeed, these lie at the intersection of \mathcal{F}_{ij} , \mathcal{F}_{ik} and \mathcal{F}_{jk} so form an $(l-3)$ -dimensional space which meets the two-dimensional \mathcal{M}^* in a point; it is such points that are labelled ijk in the figures and which mark the meeting of regions \mathcal{M}_i , \mathcal{M}_j and \mathcal{M}_k . Thus three boundary lines lines must

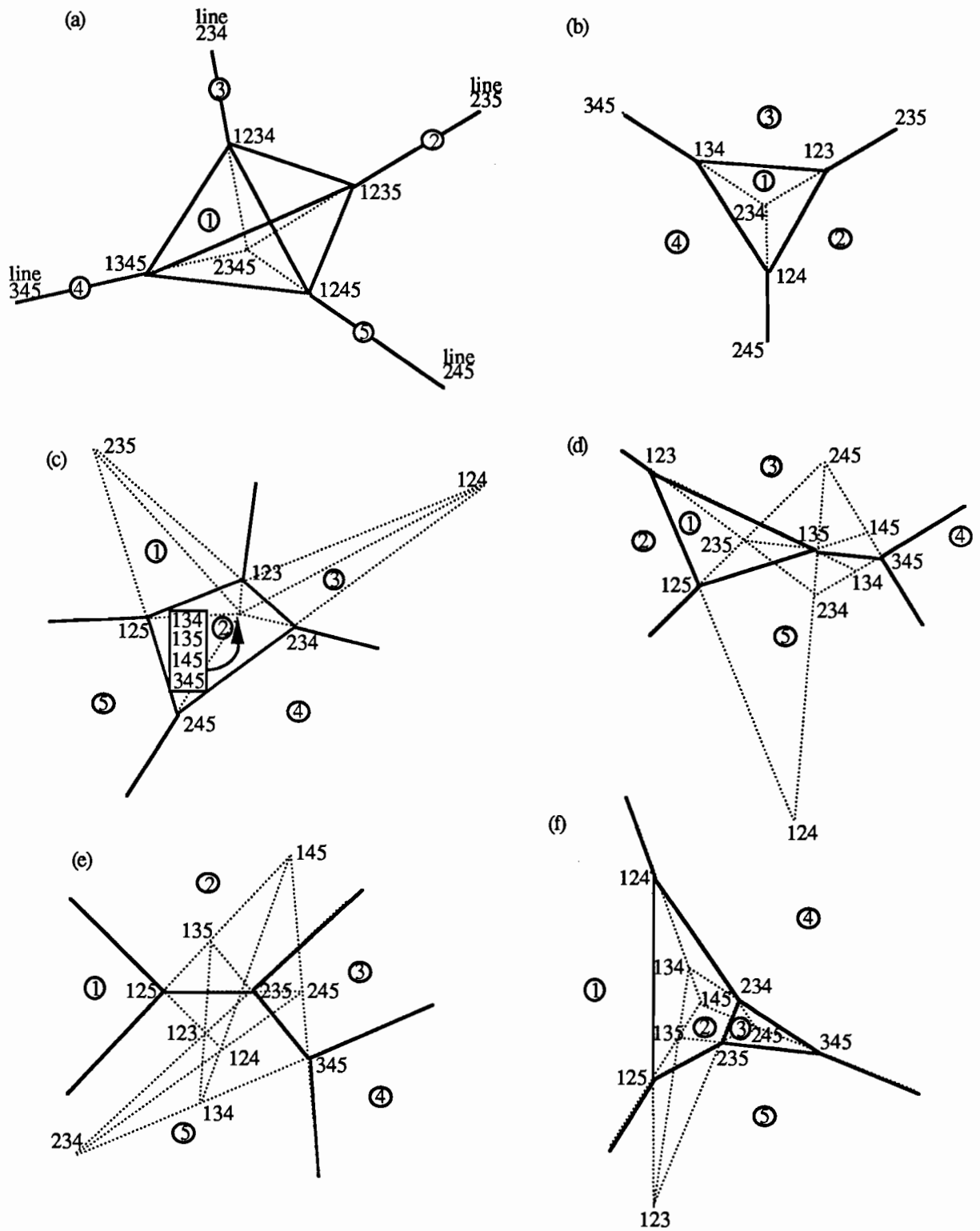


Figure 7. Additional topologies that may be obtained from two-dimensional slices of a four-dimensional simplex given by five category-level points.

meet at each triplet and the non-virtual part of each boundary is terminated by a pair of triplets that share the suffix-pair that label the regions that are separated. When there is only one non-virtual point on a line, the non-virtual boundary is completed by constructing an arm starting at the non-virtual point and extending it away from the virtual points on the

same boundary. In exceptional circumstances, two triplets ijk and ijl (virtual or non-virtual) adjacent on a line may coincide, in which case they become a point labelled $ijkl$ at the intersection of four regions, and so also coincide with ikl and jkl . Such special cases arise from a slice through a line contained in the space containing the c-centre C_{ijkl} and the lower-dimensional c-centres involving i,j,k and l ; clearly, higher dimensional analogues inducing even more concurrent regions may exist. In another exceptional circumstance \mathcal{M}^* may lie entirely within the $(l-2)$ -dimensional normal-space, in which case the order-three c-centres involving i and j meet \mathcal{M}^* in lines and the order-four c-centres meet \mathcal{M}^* in points; the regions \mathcal{M}_i and \mathcal{M}_j are then not separable for prediction. For example, in Figure 7(b) if the vertices of the triangle are labelled 1235, 1245 and 1345, then the central region is predicted as either \mathcal{M}_1 or \mathcal{M}_5 .

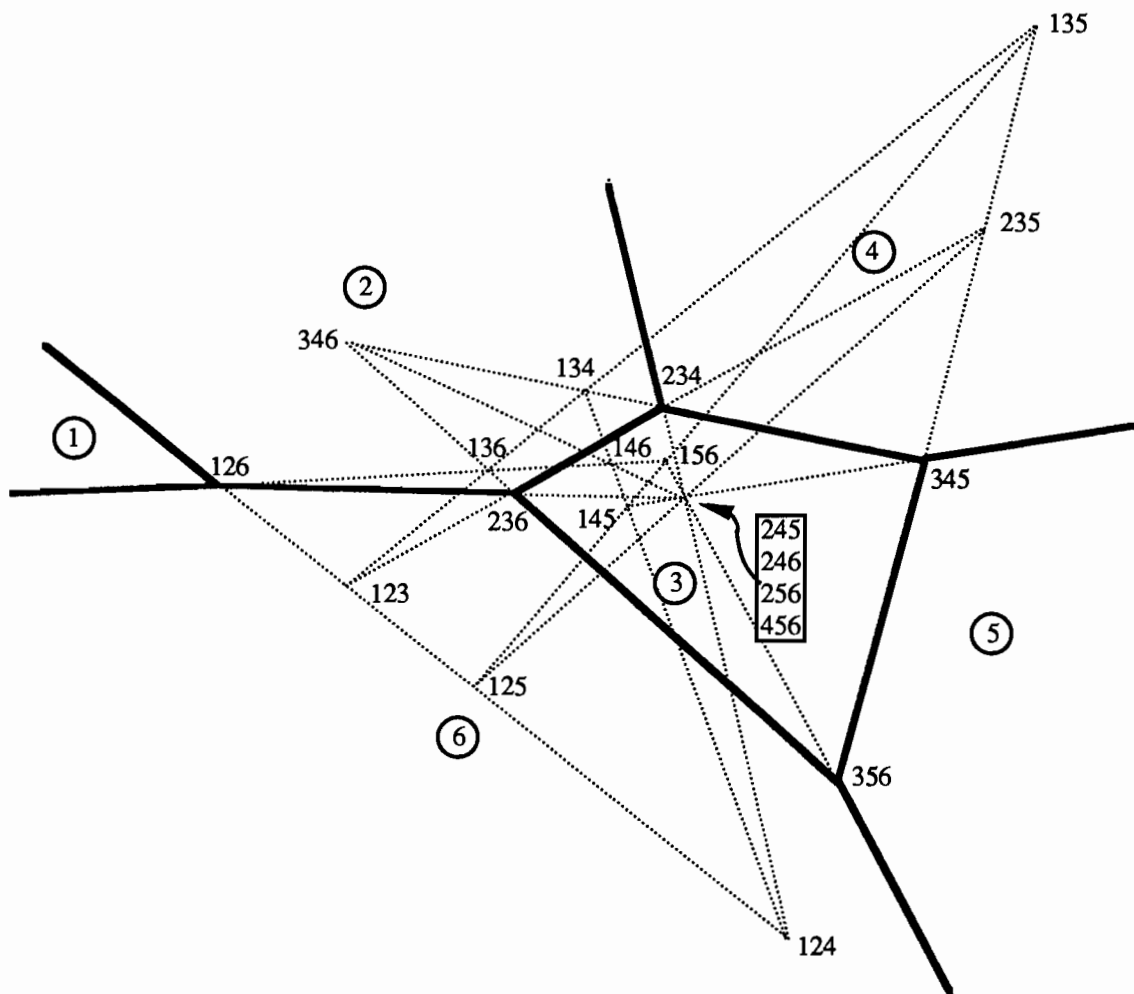


Figure 8. A two-dimensional slice of a five-dimensional simplex (six category-levels).

All triplets sharing the same pair of suffices are collinear in sets of four. Similarly to Figure 7(c), the quadrilateral contains a common point that bears the four distinct triplets marked. This quadrilateral arises from a slice through the tetrahedron with c-centre C_{2456} , and CC_{2456} meets the plane shown at the common point.

An example of the extension to five dimensions is illustrated in Figure 8; a clear connection with Finite Geometries is noted. The general form of the cut of an $(l-1)$ -dimensional simplex by a two-dimensional plane \mathcal{M}^* may now be stated:

- (1) The prediction regions in \mathcal{M}^* are convex and bounded by straight lines. Some, but not all, regions may be closed.
- (2) Each straight line contains $l-2$ triplets, most of which are virtual points, which share the two suffices labelling the regions separated by the lines.
- (3) The non-virtual triplet-points form the vertices of the polygons bounding the prediction-regions. Boundaries are formed by joining all pairs of non-virtual points that share a pair of suffices.
- (4) Special provision has to be made for boundaries that are outer arms containing only a single non-virtual point. These boundaries are at the end of a line and may be constructed by moving from the non-virtual point away from the remaining $l-3$ virtual points.
- (5) All triplets, virtual or not, are at the meeting points of three lines and so are at the join of three regions. For virtual points, these three regions are dominated by some fourth region.
- (6) Closed regions are special, in the sense that the set of lines, one from each vertex of the region, that are not parts of the region's boundary, must be concurrent at an internal point of the region. For regions bounded by more than three lines, such concurrent points represent higher-order c-centres and the superposition of several sets of triplets.
- (7) Exceptionally, higher-order c-centres may generate non-virtual points, in which case they form a vertex that is at the join of more than three prediction-regions.

3. Properties of Circumcentres

The above requires the c-centres of subsets of category-level coordinates given by the l vertices C_1, C_2, \dots, C_l of a simplex. In this paper, only the c-centres of sets of three points are required but for approximations in more than two dimensions, higher order c-centres are needed, so the following results are presented in their general form. Let the rows of \mathbf{Z} give the coordinates of the simplex. Gower (1991) gives a general method for calculating these coordinates. With the extended matching coefficient (Gower, 1991), the distances between the CLPs are all equal and the simplex is regular and then the c-centre coincides with the centroid of the CLPs and are very easily constructed. With other choices of distance, such as the chi-squared distance of multiple correspondence analysis, the CLPs will form an irregular simplex. The squared-distances c_{hk}^2 between each pair of category-level points may be calculated from \mathbf{Z} and formed into the $l \times l$ matrix $\mathbf{C} = \{-\frac{1}{2}c_{hk}^2\}$. Then Gower (1982) shows that the generalised c-centre of the coordinates \mathbf{Z} is given by $\mathbf{s}'\mathbf{Z}$, where $\mathbf{s} = \mathbf{C}^{-1}\mathbf{e}/(\mathbf{e}'\mathbf{C}^{-1}\mathbf{e})$. The neighbour-regions remain convex cones but are more complicated than for the extended matching coefficient.

In the following we shall write c_i for the i th column of C^{-1} , $c_i = e'c_i$ for the sum of the elements of the i th column, and c_{ij} for the i th diagonal element of C^{-1} . To calculate the c-centre of a face of the simplex, say the face opposite C_1 , first partition C as indicated in (5):

$$C^{-1} = \begin{pmatrix} 0 & \mathbf{b}' \\ \mathbf{b} & \mathbf{B} \end{pmatrix}^{-1} = \begin{pmatrix} \beta & -\beta\mathbf{b}'\mathbf{B}^{-1} \\ -\beta\mathbf{B}^{-1}\mathbf{b} & \mathbf{B}^{-1} + \beta\mathbf{B}^{-1}\mathbf{b}\mathbf{b}'\mathbf{B}^{-1} \end{pmatrix} \quad (5)$$

where $\beta = -(\mathbf{b}'\mathbf{B}^{-1}\mathbf{b})^{-1}$. We require $\mathbf{t} = \mathbf{T}e/(e'\mathbf{T}e)$, the centering vector for the c-centre of the face opposite C_1 , where $\mathbf{T} = \begin{pmatrix} 0 & \mathbf{0}' \\ \mathbf{0} & \mathbf{B}^{-1} \end{pmatrix}$. From (5) we have:

$$\mathbf{T} = \begin{pmatrix} 0 & \mathbf{0}' \\ \mathbf{0} & \mathbf{B}^{-1} \end{pmatrix} = C^{-1} - (c_1c_1'/c_{11}). \quad (6)$$

$$\text{Hence } \mathbf{t} = \frac{C^{-1}e - \frac{c_1}{c_{11}}c_1}{e'C^{-1}e - \frac{c_1^2}{c_{11}}}. \quad (7)$$

Thus the centering-vector for the face opposite C_1 is obtained simply by deflating the inverse of C by its first row and column and then normalising to unit total; similarly for the other faces. Clearly, by replacing C^{-1} by \mathbf{T} in (6) and (7), with corresponding changes in c_1 , these formulae may be used to obtain, by iterative deflation, the centering vectors for the c-centres of an ever-decreasing number of vertices.

Writing \mathbf{T}_1 for the matrix \mathbf{T} of (5) and \mathbf{t}_1 for the vector (6), we may similarly define \mathbf{T}_2 from which may be obtained the centering-vector \mathbf{t}_2 for the face opposite C_2 . Similarly, we may define \mathbf{T}_{12} which gives the centering-vector \mathbf{t}_{12} for the $l-2$ vertices excluding C_1 and C_2 from the full set. Repeated use of (5) shows that \mathbf{s} has the form:

$$\mathbf{s} = \gamma_1\mathbf{t}_1 + \gamma_2\mathbf{t}_2 + \gamma_{12}\mathbf{t}_{12}$$

for suitable constants γ_1 , γ_2 and γ_{12} . Writing $C_{l,12}$ for the c-centre corresponding to \mathbf{t}_{12} , this shows that the c-centres C , $C_{l,1}$, $C_{l,2}$ and $C_{l,12}$ are coplanar in two dimensions. Clearly this result remains valid for any similarly related set of c-centres, even when l refers to some subset of all of the vertices, provided that it does not contain C_1 or C_2 .

The result of the previous paragraph may be obtained, and generalised, by direct geometric argument, as follows. \mathcal{F}_{12} divides \mathcal{R}_n into two parts, one containing all the points nearer C_1 than C_2 and the other, all the points nearer C_2 than C_1 . Thus this normal has dimensionality $l-2$ and contains all c-centres sharing the pair of suffices 1,2 such as C_{12} , C_{123} , C_{1234}, \dots , $C_{123\dots l} = C$. Similarly, the space \mathcal{F}_{123} that is normal to the plane defined by C_1, C_2 and C_3 at C_{123} has dimensionality $l-3$ and contains all c-centres sharing the three suffices such as C_{123} , C_{1234}, \dots , $C_{123\dots l} = C$. Proceeding in this way we eventually find an $l - (l-2)$ plane that contains all c-centres that share $l-2$ suffices. For the $l-2$ suffices $l,1,2$, there are only the four possibilities: C , $C_{l,1}$, $C_{l,2}$ and $C_{l,12}$, as found above.

A further result which may be verified algebraically from the results of this section but which follows more directly from geometrical considerations, is that the lines $C_{\mathbf{m}}C_{\mathbf{m}.i}$ and $C_{\mathbf{m}.i}C_{\mathbf{m}.ij}$ are orthogonal for all sets of subscripts \mathbf{m} containing i and j . This follows from considering the $m-1$ vertices C_k , for all $k \approx \mathbf{m}.i$, which lie in a space \mathcal{R}_{m-2} . Their c-centre is at $C_{\mathbf{m}.i}$ and the corresponding c-sphere has radius $R_{\mathbf{m}.i}$, say. \mathcal{R}_{m-2} also contains $C_{\mathbf{m}.ij}$ for all $i \neq j \approx \mathbf{m}$. Now $C_{\mathbf{m}.i}$ is equidistant ($R_{\mathbf{m}.i}$) from all C_k . $C_{\mathbf{m}}$ also is equidistant from all C_k which lie on the c-sphere with radius $R_{\mathbf{m}}$. It follows that $R_{\mathbf{m}}^2 = R_{\mathbf{m}.i}^2 + (C_{\mathbf{m}}C_{\mathbf{m}.i})^2$ and that $C_{\mathbf{m}}$ lies on the normal to \mathcal{R}_{m-2} at $C_{\mathbf{m}.i}$. Hence, $C_{\mathbf{m}}C_{\mathbf{m}.i}$ is normal to all subspaces of \mathcal{R}_{m-2} , including the lines $C_{\mathbf{m}.i}C_{\mathbf{m}.ij}$ for all $i \neq j \approx \mathbf{m}$. Also $C_{\mathbf{m}}$ is further from C than is $C_{\mathbf{m}.i}$.

Consider vectors parallel to the lines C_1C_{12} , $C_{12}C_{123}$, $C_{123}C_{1234}$, $C_{1234}C_{12345}$, The remark at the end of the previous paragraph shows that every vector of the list is orthogonal to all those which occur to its left. Thus these vectors form an orthogonal basis for the space of the simplex, but with special reference to \mathcal{M}_1 . Clearly many other similar bases may be constructed; the rule is that in the course of constructing the list, when the next higher level c-centre is chosen, it must introduce a previously unused suffix.

4. Algorithmic Considerations

The remarks made at the end of section 2.3 form the basis of an algorithm for computing the neighbour-regions in \mathcal{L}^* but there are several details that remain to be considered. In this section, first a general outline of an algorithm is given and then some of the outstanding details in its implementation are briefly discussed. The algorithm requires an extension to the concept of joining two points. When P is a non-virtual point and Q is virtual, we require to draw a line in the direction from Q to P but starting at P ; when both points are non-virtual joining has its conventional meaning; when both points are virtual, joining is defined to have a null effect.

Algorithm

- (1) For each triplet ijk , form the space \mathcal{F}_{ijk} that is normal to $C_iC_jC_k$ at C_{ijk} . Evaluate $\mathcal{M}^* \cap \mathcal{F}_{ijk}$. This is, usually, a single point, which carries the label ijk . Determine whether or not ijk is a virtual point.
- (2) For all pairs of (ijk) and (lmn) join those that share two suffices.
- (3) Finally back-project the polygonal tessellation onto \mathcal{L} and label the regions with the category-level labels.
- (4) When $s^* < r$, complete the prediction-regions in \mathcal{L} by orthogonal extension.

Remarks

Step (1). By results given in section 3, the coordinates \mathbf{c} , say, of C_{ijk} are easily found and a pair of orthogonal vectors in the plane $C_iC_jC_k$ are given by the directions C_iC_{ij} and $C_{ij}C_{ijk}$. Suppose these two vectors are placed in the columns of a matrix $\mathbf{P} = (\mathbf{p}_1, \mathbf{p}_2)$. The matrix \mathbf{M}^* is given as described in section 1.1 as the orthogonal projection of \mathcal{L} onto \mathcal{M} . Because \mathcal{F}_{ijk} generally does not contain the origin G , the row-coordinate \mathbf{x} of any point in \mathcal{N}_{ijk} is given by $\mathbf{x} = \mathbf{c} + \mathbf{w}$, where the normality with $C_iC_jC_k$ requires that $\mathbf{w}\mathbf{P} = 0$. Also, because $\mathbf{x} \approx \mathcal{M}^*$ we have that $\mathbf{x} = \mathbf{v}\mathbf{M}^{*'}$ for some row-vector \mathbf{v} . Post-multiplying by \mathbf{P} gives $\mathbf{v}(\mathbf{M}^{*'}\mathbf{P}) = \mathbf{c}\mathbf{P}$, so determining \mathbf{v} and giving:

$$\mathbf{x} = \mathbf{c}\mathbf{P}(\mathbf{M}^{*'}\mathbf{P})^{-1}\mathbf{M}^{*'}.$$

This is the formula for determining the vertex ijk . If $\mathbf{M}^{*'}\mathbf{P}$ is singular, then \mathbf{v} is not unique and, rather than determining a single point, ijk will be a line, or even a plane, bounded by intersections with the spaces of higher order c -centres of the form $ijkl\dots$. This will occur whenever \mathcal{M}^* happens to contain an edge or lie in the surface of one of the convex cones \mathcal{M}_i . Such solutions are pathological and are not explored further here. Nevertheless, a fully robust algorithm would need to take them into consideration.

We must distinguish three kinds of vertex: those not represented, those that are virtual points and those that are non-virtual points. When the equations $\mathbf{v}(\mathbf{M}^{*'}\mathbf{P}) = \mathbf{c}\mathbf{P}$ are not consistent, there is no solution for ijk , so determining a "point at infinity" arising from a neighbour-region that is not represented in \mathcal{M}^* - see Figure 7(b) and its discussion in section 2. By definition ijk is equidistant r_{ijk} , say, from C_i, C_j and C_k . It is a virtual point if another vertex, C_l , can be found such that the distance, r_l , from ijk to C_l is less than r_{ijk} . This is easily determined by computing all r_l , for $l \neq i, j, k$.

Step (2). Recall the extended definition of joining. Its effect is as follows:

- (i) A pair of non-virtual points is joined in the conventional way
- (ii) A non-virtual point joined to a virtual point generates an arm, one end of which is unjoined to further points
- (iii) The join of any two virtual points is null and so has no visible effect.

Note that the coordinates of the ends of arms computed in Step (2) are used in Step (3).

Step (3). Back-projection is given by the formulae (4) given at the beginning of section 2 operating on all non-virtual triplets obtained in step (2). A method for labelling the back-projected regions derives from noting that every line is labelled by a pair of suffices ij , indicating the boundary between Q_i and Q_j . Either the centroid or g -circumcentre (Gower, 1985) of all the vertices on lines that share the suffix i is suggested as a suitable position for placing the label for Q_i .

References

- Bowyer, A. (1981). Computing Dirichlet tessellations. *The Computer Journal*, **24**, 162-6.
- Devijver P.A. and Diekesei M. (1985). Computing multidimensional Delauney tessellations. *Pattern Recognition Letters*, **1**, 311-6.
- Gabriel K. R. (1971) The biplot-graphic display of matrices with applications to principal components analysis. *Biometrika*, **58**, 453-67.
- Gower J.C. (1982) Euclidean distance geometry. *The Mathematical Scientist*, **7**, 1-14.
- Gower J.C. (1985) Properties of Euclidean and non-Euclidean distance matrices. *Linear Algebra and its Applications*. **67**, 81-97.
- Gower J. C. (1991) Generalised biplots. Research Report RR-91-02. Leiden: Department of Data Theory.
- Gower J.C. (1992) Biplot Geometry.
- Gower J. C. and Harding S. (1988) Non-linear biplots. *Biometrika*, **73**, 445-55.
- Sibson R. (1980) The Dirichlet tessellation as an aid to data analysis. *Scandinavian Journal of Statistics*, **7**, 14-20.
- Watson D. F. (1981) Computing the n -dimensional Delauney tessellation with applications to Voronoi polytopes. *The Computer Journal*, **24**, 167-72.