Gower, J. C. 1968. Adding a point to vector diagrams in multivariate analysis. *Biometrika.* 55 (3), pp. 582-585.

Adding a point to vector diagrams in multivariate analysis

BY J. C. GOWER

Rothamsted Experimental Station

SUMMARY

A set of n base points P_i $(i = 1, 2, \ldots, n)$, with known co-ordinates relative to orthogonal axes, and a further point P_{n+1}, with known distance from each of the base set, are given. The co-ordinates of P_{n+1} relative to the axes of the base set are found. The formula is particularly simple when the base set is referred to its principal axes, when the co-ordinates of P_{n+1} for a subset of all the axes can be calculated from the co-ordinates of the P_i in this subset only. The classical results for adding a point to a principal components or canonical variates analyses are obtained when the base set is derived using the appropriate distance functions. An example is given.

1. DERIVATION OF THE BASIC FORMULA

Gower (1966*a*) discussed how the co-ordinates of n points P_i $(i = 1, 2, \ldots, n)$ referred to principal axes can be found, given the Euclidean distances d_{ij} between all point pairs P_i and P_j; principal components and canonical variate analysis are special cases. The points are supposed to lie in an m-dimensional space; normally $m = n-1$ but in exceptional cases m can be less. The successive steps in this method may be summarized as follows:

(i) define the matrix $\mathbf{A}$ whose elements are $-\frac{1}{2}d_{ij}^2$;

(ii) define $\mathbf{B}$ with elements $b_{ij} = a_{ij} - a_{i.} - a_{.j} + a_{..}$;

(iii) find the m non-zero latent roots $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m)$ of $\mathbf{B}$ and corresponding vectors $\mathbf{X}$ standardized so that $\mathbf{X}'\mathbf{X} = \mathbf{\Lambda}$. Thus $\mathbf{BX} = \mathbf{X\Lambda}$ and

$$\mathbf{XX}' = \mathbf{B}. \tag{1}$$

The ith row of $\mathbf{X}$ gives the co-ordinates $(x_{i1}, x_{i2}, \ldots, x_{im})$ of P_i; the points are centred so that the origin is at the centroid G, i.e. the column sums of $\mathbf{X}$ are zero. When $\mathbf{B}$ is not positive semi-definite some values of λ_i will be negative and all co-ordinate values along the corresponding axes will have imaginary values. In this case the distances cannot be Euclidean but the results given here remain valid if we accept imaginary values in the usual distance formula $d_{ij}^2 = (x_{i1} - x_{j1})^2 + \ldots + (x_{im} - x_{jm})^2$.

The ith diagonal element of $\mathbf{XX}'$ may be written $x_{i1}^2 + \ldots + x_{im}^2$; this is the square of the distance d_i of P_i from the centroid, and therefore from (1)

$$b_{ii} = d_i^2 \tag{2}$$

a result which will be needed later. The geometrical interpretation of the off-diagonal elements of $\mathbf{B}$ is that $b_{ij} = d_i d_j \cos\theta_{ij}$, where θ_{ij} is the angle subtended by the line P_iP_j at the centroid, but this result will not be needed here.

When a further point becomes available, a method is needed for finding its co-ordinates

$$P_{n+1}(x_{n+1,1}, x_{n+1,2}, \ldots, x_{n+1,m}, x_{n+1,m+1})$$

relative to the axes used for the P_i $(i \leqslant n)$, allowing for an $(m+1)$th dimension which may be needed to represent exactly all the new given distances $d_{i,n+1}$ $(i = 1, 2, \ldots, n)$. The co-ordinates can be found by solving the n quadratic equations

$$d_{i,n+1}^2 = \sum_{k=1}^{m+1} (x_{n+1,k} - x_{i,k})^2 \quad (i = 1, 2, \ldots, n) \tag{3}$$

where $x_{i,m+1} = 0$ when $i \neq n+1$.

Before showing how the equations (3) can be conveniently solved, a few remarks are appropriate. In principal component analyses the point P_{n+1} will not need an extra dimension for its representation, unless m happens to be less than the total number of variates in the analysis, and the distances $d_{i,n+1}$ can be readily computed from the data. Because $x_{n+1,k}$ is the orthogonal projection of P_{n+1} onto the kth axis, the solution of (3) using these distances must agree with the classical method for adding a point in a principal components analysis; similar remarks apply to canonical variate analysis. This result, obvious geometrically, is verified below algebraically. Although m will always be less than n so that some of the n equations in (3) must be redundant, it is evident from the geometrical derivation that these equations are consistent. For example, if $n = 10$ and $m = 1$ so that the base set consists of 10 collinear points, P_{11} is fixed by two distances, say $d_{1,11}$ and $d_{2,11}$, but the remaining 8 distances will be consistent with these two. Even in the worst case there are enough equations to find all $m+1$ co-ordinates of P_{n+1}, for then the n base points will require $n-1$ dimensions for their representation so that $m+1 \leqslant n$.

To solve (3), the equations will first be written in the alternative form

$$d_{i,n+1}^2 = d_{n+1}^2 + d_i^2 - 2 \sum_{k=1}^{m} x_{ik} x_{n+1,k} \quad (i = 1, 2, \ldots, n). \tag{4}$$

Because of the centring, the cross-product term in (4) vanishes when these equations are summed over i. Thus

$$\sum_{i=1}^{n} d_{i,n+1}^2 = n d_{n+1}^2 + \sum_{i=1}^{n} d_i^2, \tag{5}$$

which can be used to substitute for d_{n+1}^2 in (4), giving after a little rearrangement

$$2 \sum_{k=1}^{m} x_{ik} x_{n+1,k} = d_i^2 - d_{i,n+1}^2 - \frac{1}{n} \sum_{i=1}^{n} (d_i^2 - d_{i,n+1}^2). \tag{6}$$

This is a set of n linear equations in the m unknowns $x_{n+1,k}$ $(k = 1, 2, \ldots, m)$. These equations are consistent because they have been derived by linear operations from the set of consistent equations (3). As all the distances on the right-hand side of (6) are known, the first m equations may be solved to find the required values, but a more symmetric method is convenient. First putting (6) into matrix form by defining $\mathbf{x}'$ to be the $1 \times m$ vector $(x_{n+1,1}, x_{n+1,2}, \ldots, x_{n+1,m})$, $\mathbf{d}$ to be the $n \times 1$ vector whose ith element is $d_i^2 - d_{i,n+1}^2$ and $\mathbf{U}$ to be the $n \times n$ matrix all of whose elements are units, we have that

$$2\mathbf{X}\mathbf{x} = \mathbf{d} - \frac{1}{n}\mathbf{U}\mathbf{d}. \tag{7}$$

Hence on pre-multiplying both sides of (7) by $\mathbf{X}'$, and noting that $\mathbf{X}'\mathbf{U} = 0$ because of the centring, and inverting, it follows that

$$\mathbf{x} = \tfrac{1}{2}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{d}. \tag{8}$$

This is a symmetric form of the solution to (7) and gives the co-ordinates of P_{n+1} in the first m dimensions, the value of $x_{n+1,m+1}$ is conveniently obtained by calculating d_{n+1} from (5) and using

$$x_{n+1,m+1}^2 = d_{n+1}^2 - \sum_{k=1}^{m} x_{n+1,k}^2. \tag{9}$$

Thus the first m co-ordinates are uniquely defined but the $(m+1)$th is determined in value but not sign.

So far the principal axis property of $\mathbf{x}$ has not been used; the results (8) and (9) are applicable to any centred base set $\mathbf{X}$ referred to orthogonal axes. When $\mathbf{X}$ is referred to principal axes (8) becomes

$$\mathbf{x} = \tfrac{1}{2}\mathbf{\Lambda}^{-1}\mathbf{X}'\mathbf{d}. \tag{10}$$

Because $\mathbf{\Lambda}$ is diagonal, the value of $x_{n+1,k}$ is determined solely from λ_k, the kth column of $\mathbf{X}$ and the elements of $\mathbf{d}$. The quantities d_i^2 occurring in $\mathbf{d}$ are diagonal elements of $\mathbf{X}\mathbf{X}'$ and their evaluation would seem to need all the columns of $\mathbf{X}$ but it was shown in (1) that $\mathbf{X}\mathbf{X}' = \mathbf{B}$, a matrix whose elements are simply computed and which is in any case needed to compute even only one column of $\mathbf{X}$. Thus if the analysis leading to (1) is used to give a representation of the base-set of points P_i in r $(< m)$ dimensions, e.g. by ignoring axes with small λ_k, (10) can still be used to find the co-ordinates of the projections of P_{n+1} on to the reduced space using only r columns of $\mathbf{X}$, the known true distances d_i of $P_i (i = 1, 2, \ldots, n)$ from their centroid, obtained from the matrix $\mathbf{B}$, and the known true distances $d_{i,n+1}$. The residual distance $d_{n+1,r}$ of P_{n+1} from the r-dimensional plane is then given by a modification of equation (9) and is

$$d_{n+1,r}^2 = d_{n+1}^2 - \sum_{k=1}^{r} x_{n+1,k}^2. \tag{11}$$

When the co-ordinates of the base-set are not referred to principal axes formula (8) must be used, and this requires all m columns of $\mathbf{X}$ to compute $(\mathbf{X}'\mathbf{X})^{-1}$.

2. Principal component and canonical variate analysis

A simple geometrical argument has shown that (10) must reduce to the desired results when adding a point in a principal components or canonical variates analysis. This is not obvious from an inspection of (10) but is readily verified algebraically.

In a canonical variates analysis on n populations with v measured variates and common dispersion matrix $\mathbf{W}$, the means for the ith population will be written as the $1 \times v$ vector $\mathbf{g}_i$ $(i = 1, 2, \ldots, n)$. The $n \times v$ matrix $\mathbf{G}$ of all the population means has rows $\mathbf{g}_i$ and is supposed centred so that the column sums, ignoring any differences in population sizes, are zero. The latent roots $\mathbf{\Lambda}$ and vectors $\mathbf{L}$ satisfying

$$\mathbf{G'GL} = \mathbf{WL\Lambda} \tag{12}$$

can be found and the population means referred to canonical variate axes are defined by

$$\mathbf{X} = \mathbf{GL}, \tag{13}$$

where the vectors are normalized so that $\mathbf{L'WL} = \mathbf{I}$. Reasons for sometimes using $\mathbf{G'G}$ rather than the usual weighted between-population dispersion matrix were discussed by Gower (1966*b*). The co-ordinates $\mathbf{X}$ of the base set are defined by (13) and the usual formula giving the co-ordinates $\mathbf{x}(v \times 1)$ referred to the canonical axes of a new sample $\mathbf{g}(1 \times v)$ referred to the original variate axes is

$$\mathbf{x} = \mathbf{L'g'}. \tag{14}$$

Mahalanobis's D^2 is the appropriate distance for this type of analysis and therefore

$$d^2_{n+1,i} = (\mathbf{g} - \mathbf{g}_i)\,\mathbf{W}^{-1}(\mathbf{g} - \mathbf{g}_i)' \quad \text{and} \quad d^2_i = \mathbf{g}_i\mathbf{W}^{-1}\mathbf{g}'_i. \tag{15}$$

Therefore $$d^2_i - d^2_{n+1,i} = 2\mathbf{g}_i\mathbf{W}^{-1}\mathbf{g}' - \mathbf{g}\mathbf{W}^{-1}\mathbf{g}' \quad \text{and} \quad \mathbf{d} = 2\mathbf{G}\mathbf{W}^{-1}\mathbf{g}' - \mathbf{g}\mathbf{W}^{-1}\mathbf{g}'\mathbf{1}, \tag{16}$$

where $\mathbf{1}'$ is an $n \times 1$ vector of units. Insertion of these results in (10) gives

$$\tfrac{1}{2}\mathbf{\Lambda}^{-1}\mathbf{X}'\mathbf{d} = \tfrac{1}{2}\mathbf{\Lambda}^{-1}\mathbf{L}'\mathbf{G}'(2\mathbf{G}\mathbf{W}^{-1}\mathbf{g}' - \mathbf{g}\mathbf{W}^{-1}\mathbf{g}'\mathbf{1}). \tag{17}$$

The second term on the right-hand side of (17) vanishes because $\mathbf{G'1} = 0$ so that

$$\tfrac{1}{2}\mathbf{\Lambda}^{-1}\mathbf{x}'\mathbf{d} = \mathbf{\Lambda}^{-1}\mathbf{L}'\mathbf{G}'\mathbf{G}\mathbf{W}^{-1}\mathbf{g}',$$

which simplifies by using the transpose of (12) to

$$\tfrac{1}{2}\mathbf{\Lambda}^{-1}\mathbf{X}'\mathbf{d} = \mathbf{\Lambda}^{-1}\mathbf{\Lambda}\mathbf{L}'\mathbf{W}\mathbf{W}^{-1}\mathbf{g}' = \mathbf{L}'\mathbf{g}', \tag{18}$$

agreeing with (14) as required.

The verification for principal components is almost identical, but with $\mathbf{W}$ replaced by $\mathbf{I}$, and $\mathbf{G}$ interpreted as a $n \times v$ data matrix derived from observations on v variates for a multivariate sample of size n.

3. Example

Columns 2, 3 and 4 of Table 1 were derived from a table giving the distances d_{ij} between every pair of eleven British cities, using the method described by Gower (1966*a*) and outlined here in the discussion leading to (1) and (2). These constitute the calculations needed to find the co-ordinates of the base-set, here in two dimensions, and remain fixed when positioning a new city. The third, fourth and fifth roots are 6878, 2338 and 288, all small compared with the first two, given in Table 1, and the five remaining roots are negative with a total value of -5966, less in modulus than λ_3. Therefore, although an exact reproduction of the given distances is impossible in two dimensions, or in any number of real dimensions, these co-ordinates give quite a good Euclidean representation of the relative positions of the cities; the road distances as reproduced are, on the average, about 1·13 times the direct distances.

The squared distances $d^2_{n+1,i}$ of a further city, Birmingham, from each city of the base set are given in column 5 and the elements of the vector $\mathbf{d}$, found by subtracting column 5 from 4, are given in column 6. The latent roots λ_1 and λ_2 found in the base set analysis are required and it can be checked from Table 1 that

$$\lambda_1 = \sum_{i=1}^{11} x^2_{i1} \quad \text{and} \quad \lambda_2 = \sum_{i=1}^{11} x^2_{i2}.$$

The first co-ordinate $x_{12,1}$ of Birmingham is found as the sum of products of columns (2) and (6) divided by $2\lambda_1$; this gives

$$x_{12,1} = -69{,}6384{\cdot}26/(2 \times 177{,}738) = -2{\cdot}0.$$

Similarly, $$x_{12,2} = -2{,}241{,}066/(2 \times 29{,}178) = -38{\cdot}4.$$

The position $(-2{\cdot}0, -38{\cdot}4)$ places Birmingham very well. Its distance d_{12} from the centroid of the base set is found from equation (5) as

$$11d_{12}^2 = 232{,}505 - 210{,}653{\cdot}7 = 21{,}851{\cdot}3.$$

Table 1. *Quantities needed to determine the position of Birmingham*

1 City, i	2 x_{i1} (miles)	3 x_{i2} (miles)	4 $d_i^2 = b_{ii}$	5 $d_{n+1,i}^2$ (Birmingham)	6 (4)–(5) = d
Brighton	−140·7	9·3	21,268·7	25,600	−4,331·3
Bristol	−72·9	−92·6	14,424·3	7,744	6,680·3
Cambridge	−39·8	52·6	3,856·2	10,000	−6,143·8
Edinburgh	283·6	−14·9	80,785·9	82,944	−2,158·1
London	−88·8	20·3	8,492·4	12,100	−3,607·6
Manchester	78·3	−31·9	8,333·3	6,400	1,933·3
Newcastle	184·3	12·1	34,902·0	40,401	−5,499·0
Norwich	−43·9	113·4	15,632·3	24,336	−8,703·7
Nottingham	28·7	5·5	439·7	2,500	−2,060·3
Oxford	−62·7	−24·1	4,483·8	4,096	387·8
Southampton	−126·1	−49·7	18,035·1	16,384	1,651·1
Latent roots	177,738	29,178	—	—	—
Totals	—	—	210,653·7	232,505	—

Columns 1, 2, 3 and 4 are given by the calculations for the base-set and are determined in a preliminary analysis. Column 5 is given and 6 is the difference between 4 and 5.

Thus $d_{12}^2 = 1986{\cdot}5$, i.e. minus the mean of column 6 and the residual distance $d_{12,r}$ of Birmingham from the plane of the base set is found from equation (11) to be 22·5 miles. It must be remembered that this residual has real and imaginary components so that $d_{12,r}^2$ may become small because of the possible cancellation of large positive and negative components; this would not be so if the original distances had been Euclidean giving no negative latent roots. In this example the root-mean-square residual of the base set in the direction of the third dimension is $(6878/11)^{\frac{1}{2}} = 25{\cdot}0$ miles and the total root-mean-square residual out of the fitted plane, using positive and negative roots, is $(3538/11)^{\frac{1}{2}} = 17{\cdot}9$ miles, both agreeing well with $d_{12,r}$. However, a critical examination of this residual would require the real and imaginary components, separately.

A glance at an atlas reveals that Birmingham is well within the region covered by the base set and the usual warnings apply to extrapolation outside this region; the position of John O'Groats would be poorly determined with the base set chosen here.

References

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

Gower, J. C. (1966*b*). A *Q*-technique for the calculation of canonical variates. *Biometrika* **53**, 588–9.

[*Received February* 1968. *Revised April* 1968]