STATISTICS IN MEDICINE, VOL. 9, 469–485 (1990)

BIPLOTS IN BIOMEDICAL RESEARCH

K. RUBEN GABRIEL AND CHARLES L. ODOROFF*

Department of Statistics and Division of Biostatistics, University of Rochester, Rochester, NY 14627, U.S.A.

SUMMARY

The biplot is a graphical display of multivariate data. A number of examples from biomedical research illustrate its use. Biplots allow for inspection of data preliminary to formal analysis, and can follow an analysis by the graphical inspection of residuals. The simplicity and intuitive appeal of these displays is stressed. An appendix indicates their method of construction and the software for producing them.

INTRODUCTION: STORKS AND THE BIRTH RATE

A biplot[1] is a graphical display of multivariate data.[2–5] As a scatterplot shows the joint distribution of two variables, so a biplot shows it for three or more variables. Exact representation of three variables needs three dimensions (3-D), of four variables four dimensions (4-D), and so on, but approximate representation by means of biplots is possible in the plane. Biplots usually reveal the main features of higher dimensional data even when they use only two dimensions (2-D). Closer approximations obtain with 3-D biplots, but their effective display requires real time computer graphics,[6,7] whereas a 2-D biplot plots simply on a sheet of paper.

An example introduces the biplot. Figure 1 displays 1953–77 Danish data from Table I on the frequency of nesting storks, the birth rate of humans and the per capita consumption of electricity.[8] The units of observation are the individual years, and these are plotted and labelled by the dates, that is 53, 54, . . . , 77. The three variables are represented by linear axes with scales used in the same way as in an ordinary scatterplot. (In order not to clutter the display, only the lowest and highest values on each axis are labelled explicitly; for example, only the values 12 and 20 are printed on the birth rate axis. Other values can be interpolated.) We approximate unit i's observation on variable j by projecting the mark for unit i perpendicularly onto the axis for variable j and reading the value on the scale. For example, for year 1953 we project the 53 marker onto the stork scale to read a value of about 180, onto the birth rate scale for a value around 17·9, and onto the electricity scale for one of about 720. These projected scale values are approximations to the true values 177, 17·9 and 701 in Table I. They are not exact – even to the accuracy of graphical representation – because it is not usually possible to represent more than two variables exactly in the plane. But the fit is close enough to yield a good grasp of the features of this data set. Indeed, in this example the sum of squares of the deviations from the biplot plane (that is, the inaccuracies of the approximations) is only 1·1 per cent of the sum of squares of the three variables' deviations from their means.

* Deceased.

0277–6715/90/050469–17$08.50

Received January 1989
Revised October 1989

Table 1. Data from Denmark 1953–1977[8]

	Nesting of storks	Birth rate	Consumption of electricity
1953	177	17·9	701
1954	210	17·3	779
1955	217	17·3	849
1956	214	17·2	867
1957	199	16·8	909
1958	186	16·5	962
1959	165	16·3	1042
1960	145	16·6	1153
1961	135	16·6	1270
1962	154	16·7	1435
1963	121	17·6	1584
1964	111	17·7	1725
1965	106	18·0	1916
1966	101	18·4	2069
1967	87	16·8	2227
1968	73	15·3	2438
1969	65	14·6	2709
1970	60	14·4	3197
1971	54	15·2	3358
1972	51	15·2	3667
1973	45	14·3	3763
1974	40	14·1	3701
1975	39	14·2	3871
1976	37	12·9	4289
1977	35	12·2	4533
Scales for biplotting	1/100	1	1/1000
Means (scaled)	1·135	16·004	2·201
SDs (scaled)	0·634	1·665	1·254
Correlations			
with storks	1·000	0·722	− 0·940
with births	0·722	1·000	− 0·855
with electricity	− 0·940	− 0·855	1·000

Consideration of the features of these data demonstrates what one can see on a biplot. The projections of the years' markers onto the stork axis show that the number of stork nestings was highest in the 1950s, then fell rather consistently to the end of the period, except for fairly stationary periods during 1966–68 and during 1972–74. Next, the projections of the same markers onto the birth rate axis show a decline from 1953 to 1959, a rather strong recovery to 1966, and a subsequent almost continuous decline to 1977 (except for a mild recovery in 1971 and 1972). Evidently, the development of these two series is similar in that both decline with time, and therefore they are highly correlated. This may encourage those who still want to explain the declining birth rate by a dearth of storks. (That is the European parallel to the American cabbage patch theory of the origin of babies.) However, the projections of the years' points onto the electricity axis exhibit an almost consistent increase with time, with a trend that is slow at first, then picks up speed, and finally slows down slightly in the early 1970s. One could ask what made the consumption of electricity rise: the dearth of stork nestings or the scarcity of babies? Or vice

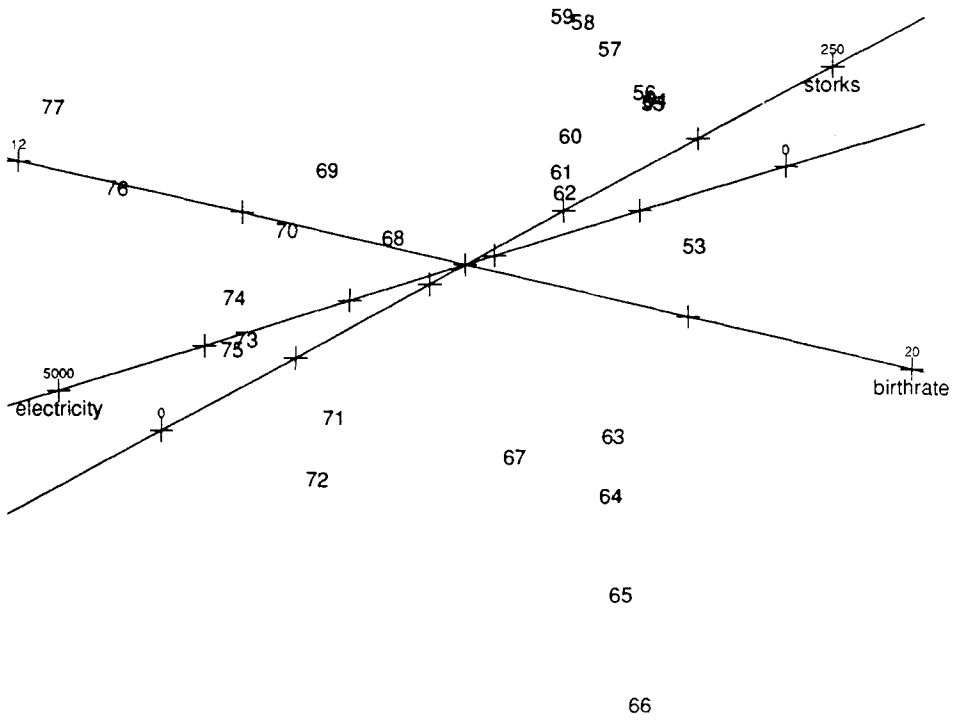

Figure 1. Denmark: storks, electricity and births, scatter and axes. Goodness of fit 99·45 per cent, centred on means

versa? Whatever 'scientific' interpretation one chooses, these data illustrate what a biplot is and how one reads it.

It is clear that with any two correlated variables, their biplot axes must go in similar directions; for the markers with high (low) projections onto one axis will mostly also have high (low) projections onto the other axis. Similarly, with negatively correlated variables such as electricity and births, the high (low) projections onto one axis are generally associated with low (high) projections onto the other, and so the axes must go in opposite directions. Hence, an acute angle between the positive directions represents highly correlated variables, an obtuse angle negatively correlated variables, and a right angle uncorrelated variables. In the present example, the stork and birth series are positively correlated with one another and negatively with the series for electricity. The biplot reflects this by the acute angle between the directions for storks and for births, and by the obtuse angles of these directions with that for electricity.

A useful convention, as in Figure 1, is to construct biplots so that the correlations between variables are represented approximately by means of the cosines of the angles between the corresponding axes. Thus the angle between the axes for births and for storks is 42°; the correlation between these two variables is 0·72, quite similar to $\cos(42°) = 0·74$. Also, the angle between the axes for electricity or for storks is 168° (relative to the positive directions of each); $\cos(168°) = -0·98$, close to the correlation of $-0·94$ between these two variables.

Another useful convention used in this display is to have all the axes meet at the centroid, which is the mark for the means of all the variables. Thus, for a unit located at this point, its value on each variable is the mean of that variable. In this example, 1968 appears closest to an 'average' year on all three variables.

Biplots usually do not show the scales for the variables explicitly. Instead, arrows from the origin show the positive directions of the variables' axes, and the lengths of the arrows approximate the standard deviations of the variables. Also, the angle between any two arrows (which is the same as that between the positive directions of the axes) approximates the arc cosine of the correlation between those variables, as illustrated above. Such biplots are called column metric preserving, or CMP for short.[1,3,5] Mathematical details of their construction, as well as available software, are described in the appendix.

Figure 2 shows the CMP biplot of the data of Table I in this point-and-arrow form. The angles are the same as in Figure 1, as they represent the correlations in the same way. The lengths of the arrows indicate that the standard deviation of the stork series was much smaller than those of electricity and of the birth rate, the latter being largest. (Note how this corresponds to the true standard deviations in Table I.)

The point-and-arrow mode of display of the biplot (Figure 2) is rather less cluttered than the point-and-axis mode (Figure 1). The former makes it easier to assess the general features of the data, but the latter is better for obtaining approximate data values. Clearly, these are not two different biplots, but merely two modes of display of the same biplot, one more suitable for detailed inspection of the data and the other more suitable for appraisal of its general features. For large data sets, the point-and-axis mode may be impractical because too many data make detailed inspection difficult.

COMPARING TREATMENT AND CONTROL GRAPHICALLY: ANALGESIC TRIALS

A second example of a CMP biplot, Figure 3, displays data from an analgesic trial in which 20 surgical patients post-operatively received a single dosage of analgesic and 20 others received a placebo.[9,10] Pain scores were recorded for each patient at medication time and 0·5, 1, 2, 3, 4 and 5 hours later. These 40-unit, seven-variable data were centred on the mean pain score for each time and then biplotted.

The seven biplot arrows of Figure 3 show the joint variation of the scores at different times. For the initial time, at medication, there is a very short biplot arrow – evidently because all patients had very high pain scores at that time, so there was nearly zero variability. The arrow for 30 minutes after medication is longer, showing higher variability, and the lengths of the remaining arrows are roughly equal, which shows that variability stayed relatively constant from 1 to 5 hours after medication.

The angles between the arrows reflect the correlations between the pain scores at different times. The arrow for 1 hour is very close to that for 30 minutes, the one for 2 hours is farther away, and those for 3, 4 and 5 hours yet farther away. Generally, the angles between arrows for scores closer in time – as 1 and 2 hours, 3 and 4 hours, and so on – are more acute than the angles between arrows for scores farther apart in time, as 1 and 3 hours, 3 and 5 hours, and so on. In other words, the general pattern indicates that greater time separation gives a wider angle, which means a lower correlation. Conversely, scores closer in time have arrows with smaller angular separation, showing a higher correlation. Indeed, the scores at 30 minutes and at 1 hour after the operation were very highly correlated (the angle of 8° indicates a correlation of $\cos(8°) = +0{\cdot}99$), as were those at 4 and 5 hours. On the other hand, the scores at 30 minutes hardly correlated at all with those at 5 hours (the angle of 79° corresponds to the low correlation of $\cos(79°) = +0{\cdot}19$).

The fan pattern of the arrows of this biplot reflects a correlation pattern that makes clear clinical sense.

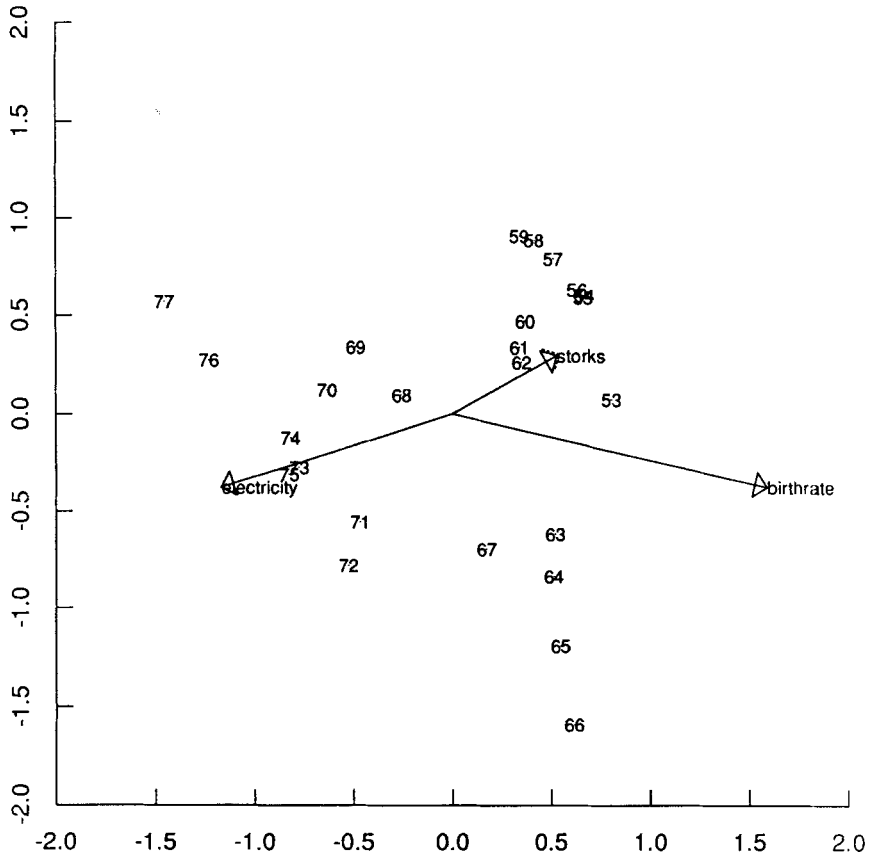

Figure 2. Denmark: CMP biplot of storks and other data. Goodness of fit 99·45 per cent, centred on means

One can also visualize combinations of variables on the biplot. For example, we represent the average of a number of variables by their average arrow (defined so that the arrowhead's coordinates are the averages of the coordinates of the individual variables' arrowheads). On Figure 3 such a construction would show that the average pain score at 3, 4 and 5 hours was similar to that for 4 hours but had slightly less variability (its arrow would be a little shorter than that for 4 hours). Another useful construct is illustrated by the difference between the pain scores at 5 hours and at 1 hour (each measured from its mean), which is represented by a biplot arrow whose length and direction are similar to the vector from the head of the arrow for 1 hour to that for 5 hours, that is approximately vertically upward.

In this example, the main interest is in comparing all patients treated with analgesics versus all those receiving placebo; there was less interest in the individual patients. The treatment is shown in Figure 3 by the labels marking each patient's location on the biplot. The treated patients' markers generally scatter more to the right and downward, whereas the placebo receivers' markers are mostly farther to the left and upward. The two groups' separate scatters are summarized in Figure 4 by means of concentration ellipses. Each of these ellipses is centred on its group's centroid and its width in any direction expresses the group's mean $\pm$ standard deviation on the variable displayed in that direction. (A concentration ellipse is the multivariate counterpart of the single variate interval of plus or minus one standard deviation from the mean.[2,10]) One

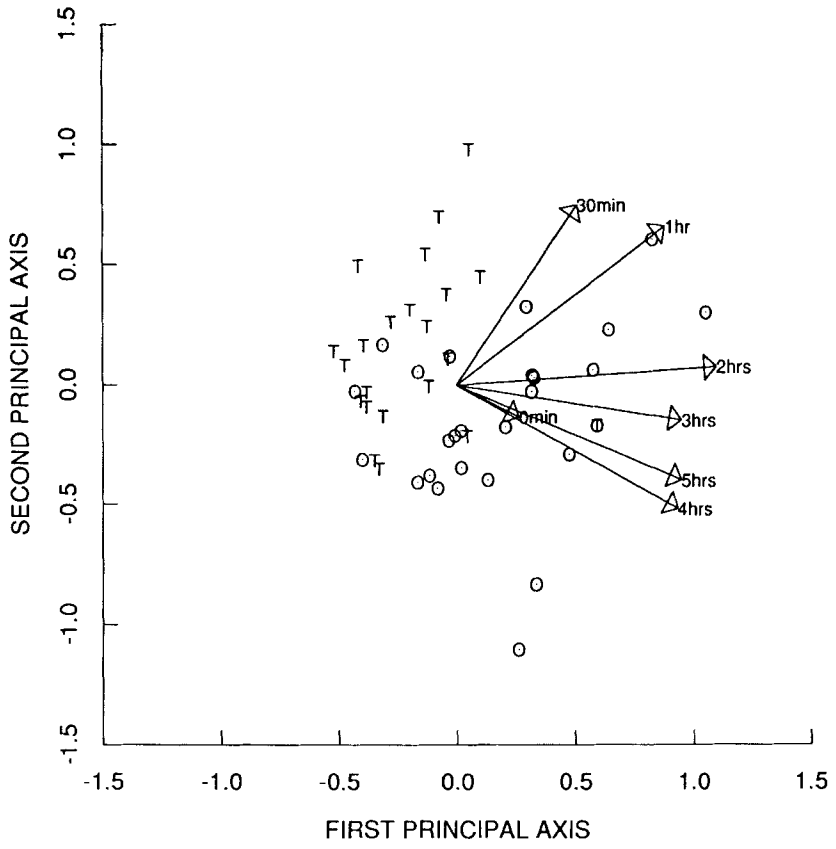

Figure 3. Analgesic trials: pain scores, scatter and arrows. T, treatment; O, placebo. Goodness of rank 2 fit is 0·734047

may project each ellipse of Figure 4, or the markers of the corresponding 20 patients of Figure 3, onto the variable axes and see how the two groups differed in pain scores. Doing so shows that the projections onto the direction of the arrows for the early scores (30 minutes and 1 hour) are much the same for the two groups, which indicates that in the short run the placebo was as 'effective' as the analgesic. It also shows that the projections onto the arrows for the later hours differentiate the groups clearly; almost all the placebo markers' projections are on the positive side of the centroid and most of those of the treated markers are on the negative side. Thus, 3 and more hours after treatment, there was a big difference between patients who had received the analgesic and those who had not, with the former's pain scores below, and the latter's above, the mean.

Again, this makes clinical sense: the placebo effect was overwhelming in the short run, within the first hour or so, but wore off later.

ON THE CONSTRUCTION OF BIPLOTS

The labels of the axes in Figures 3 and 4 indicate first and second principal axes (PCs). This is because the coordinates of a CMP biplot can be obtained[1] from a principal components analysis (PCA). The coordinates of the *i*th marker are unit *i*'s scores on the first two PCs (when normalized to unit variance). Also, the coordinates of the head of arrow *j* are variable *j*'s loadings with respect

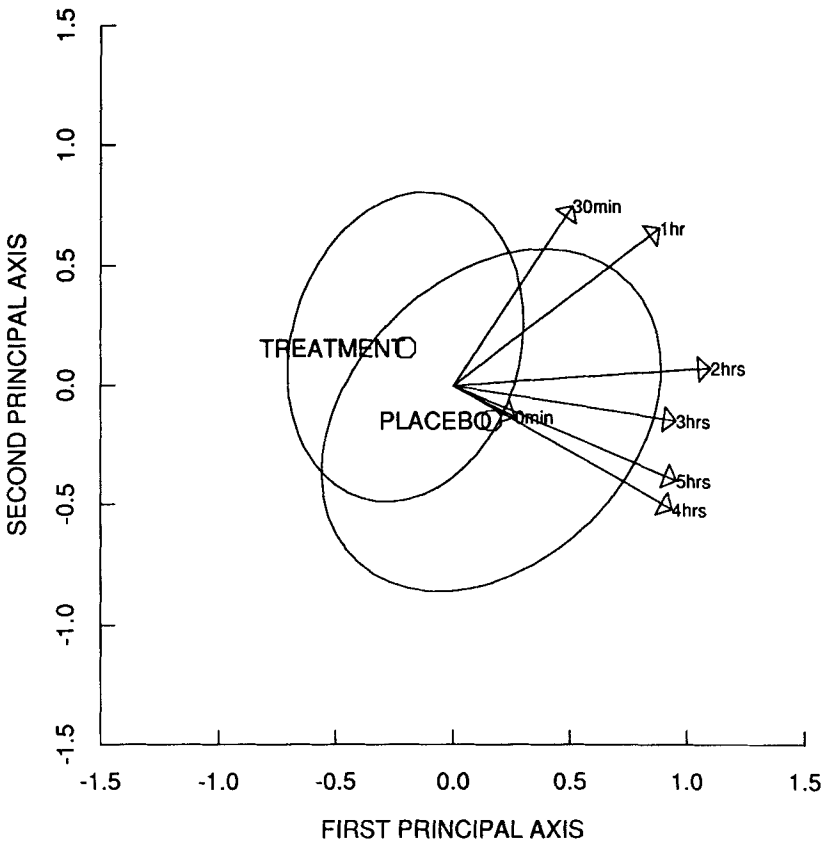

Figure 4. Analgesic trials: pain scores, concentration ellipses and arrows. Goodness of rank 2 fit is 0·734047

to the first two PCs, each weighted by the standard deviation of that component, that is by the square root of the corresponding eigenvalue. (For the exact formulas, see the appendix.)

This construction ensures a number of things which follow from the mathematics of PCA as applied to centred data, that is to observations $y_{i,j}$ of unit i on variable j corrected for that variable's mean m_j. First, the biplot gives the least squares approximation of the centred data in a plane, with a measure of its goodness of fit given by the formula

$$1 - (\text{sum of first two PCA eigenvalues})/(\text{sum of all PCA eigenvalues}).$$

(This approximation depends on the scales for presentation of the individual variables, just as in PCA. If the numbers used to represent variable j are larger than those for variable j', the fit of j will be proportionally closer than that of j'. It is the investigator's choice of scales that determines which variables are more closely approximated.) Second, the configuration of the CMP biplot arrows for the variables provides approximations to the variability and correlations of the variables, in that the length of the jth arrow approximates the standard deviation s_j of the jth variable, and the cosine of the angle between arrows j and j' approximates the correlation between variables j and j'. Third, one may visualize linear combinations of the original variables on the biplot by combinations of the vectors; for example, the biplot arrow for the difference of variables j and j' is the vector difference between arrows j and j'. Fourth, the scatter of the units'

markers reflects statistical similarities between units in that the markers of similar units will generally appear close together, and those for dissimilar units farther apart. Finally, $(y_{i,j} - m_j)/s_j$, the ith unit's standardized observation on variable j, is displayed by the coordinate of the marker for i with respect to axis j, that is by the magnitude of the intercept from the CMP biplot's origin to the foot of the perpendicular from the ith marker onto the axis through the arrow for variable j.

Evidently, the centred observation $y_{i,j} - m_j$ itself is displayed by multiplying the magnitude of that origin-to-perpendicular intercept by the length of the arrow. That allows the construction of a scale for variable j along that axis, with the origin corresponding to m_j and the head of the jth arrow corresponding to $m_j + s_j$ (as in Figure 1).

The magnitudes of the coordinates for the units' markers may differ substantially from those for the variables' arrowheads. To plot them together may therefore require the use of different scales for markers than for arrows. This does not mean, however, that one may use different scales horizontally and vertically; doing that, either for units or for arrows, would distort the biplot and destroy the properties described above.

One can also construct biplots based on other criteria of approximation,[11,12] a topic beyond the scope of the present paper. Also, if the representation of standard deviations and correlations has little interest, a construction different from the CMP may be appropriate; an illustration is given later. Apart from that contingency table example, this paper concentrates on the most generally useful form, the CMP biplot.

An investigator who is interested in the principal components of the data need only look at the biplot horizontally for the first PC and vertically for the second PC. For example, in Figure 1 the first PC corresponds closely to the electricity variable and is highly correlated (small angle) with each of the other two variables. What is the second PC? It is difficult to label it. Also, in the example of Figure 3, the horizontal direction seems to constitute an average pain score and the vertical direction a measure that differentiates between the earlier and later pain scores. Does use of these PCs provide more insight than the earlier description of the time related fan pattern of the biplot arrows?

The interpretation of a CMP biplot need not make any use of the PCA aspects of its construction. In other words, though one uses the principal axes and principal scores in constructing the biplot, they need not play any role in its interpretation.[13] For example, the above discussion of the two biplot examples made no mention of the principal axes. Indeed, the discussion would be identical if either of these biplots were rotated through any angle and/or reflected.

Whether or not to discuss the principal components explicitly, or use them only as a tool for construction, is a decision for the individual investigator. But there is something to be said for a simple description of the data directly in terms of units and variables – as done with a biplot – rather than with the introduction of mathematical constructs that themselves need to be defined in terms of the variables.[13]

With comparison of groups of units, interest is likely to focus on location of the group means and variation relative to each other and to the centroid. Strauss *et al.*[14] provided an example of using a biplot for the comparison of several groups, when they explored the planar structure of psychiatric diagnoses, among both patients and archetypes as conceptualized by psychiatrists.

In the analgesic example, interest focused on group effects (analgesic versus placebo). The biplot labels therefore indicated only the group, not the individual. By contrast, in the storks illustration interest was in the years, so the units were labelled individually. This allowed the establishment of a time trend.

Another use of the biplot often occurs in the initial inspection of a data set when one finds, near the centroid, a tight blob with almost all the units' markers, and a very small number of markers

farther out along the axes of one or two variables. That usually flags a few erroneous observations or outliers, and may lead to their exclusion or to the correction of some of their measurements. Our practice has been to separate them from the main analysis and report on them explicitly (as in a footnote) since these observations may at times have considerable scientific interest. We then construct a new biplot without them. (The inclusion of even a single highly aberrant observation could distort the entire biplot.[11])

ALZHEIMER AND CONTROL BRAIN DATA DISPLAYED IN THREE DIMENSIONS

Preliminary data from a study on Alzheimer's disease (Hamill *et al.*[15]) illustrate inspection of data by a 3-D biplot. There were 7 Alzheimer and 7 control brains, each with the following measurements on the middle frontal gyrus: counts of senile plaques and neurofibrillary tangles (P&T); counts of neurons (Neur); choline acetyltransferase activity, expressed in logarithms (Cat); size of somatostatin containing neurons (Soma); and dendritic extent (Dend). Age at death was also recorded (Age).

These data – after centring all variables on their means and scaling them to similar variances – are not adequately fitted by a planar biplot. Addition of a third dimension, however, permitted good approximation by a 3-D biplot.

A rough picture of the 3-D display is given in Figure 5(a), which shows the frontal view of this biplot and uses perspective and shading to simulate the third dimension; thus, for example, the large objects are the ones closer to the front and the smaller objects are farther behind. (For the actual data analysis we used the ANIMATE module,[7] which provided perspective, colour and motion cues but could not be reproduced on a black and white page.)

This biplot shows the dependence of the various measures. The only high correlation was of P&T with Soma. To see the structure for the remaining correlations, note that the vectors for Age, Neur and Dend are close to a plane, and that the angle between the vectors for Age and Dend is highly obtuse, which indicates strong negative correlation. The vector for Neur is in between, which shows that this variable had some positive correlation with Age and with Dend. Also, the vectors for P&T and Soma as well as Age and Cat are roughly ordered on another plane, with Age between P&T and Soma on one side and Cat on the other. Evidently, each of P&T, Soma and Cat was somewhat positively correlated with Age.

All that makes good clinical sense, but of more interest was the differentiation between Alzheimer and control brains, highlighted in this display by the difference between cubes (representing controls) and octagons (which represents the Alzheimer cases). In the original view (Figure 5(a)) there is no clear spatial separation between the two types of objects, but after rotating the biplot by 45° about the vertical axis (Figure 5(b)) the two sets of objects clearly separate from each other. The fact that such a separation occurs shows that this multivariate set of measures could discriminate between Alzheimer and other brains.

The biplot not only shows that discrimination is possible, but also reveals which variables allowed this discrimination. Figure 5(b) indicates vertical separation by the plane which was noted to be close to the vectors for Neur, Dend and Age. Hence the discrimination was on the other three variables. Indeed, the vectors that point to the right (towards the Alzheimer markers) are those for P&T and Soma, whereas the Cat vector points to the left (away from the Alzheimer markers). Again, this confirms and strengthens previous clinical knowledge about Alzheimer's as a disease of failure of multiple neuronal systems.

This example demonstrates how biplot display reveals features of multivariate data not evident from study of the individual measurements. It has leaned on technologically sophisticated methods of display, but these are rapidly becoming accessible at a reasonable cost to the user.[6,7]

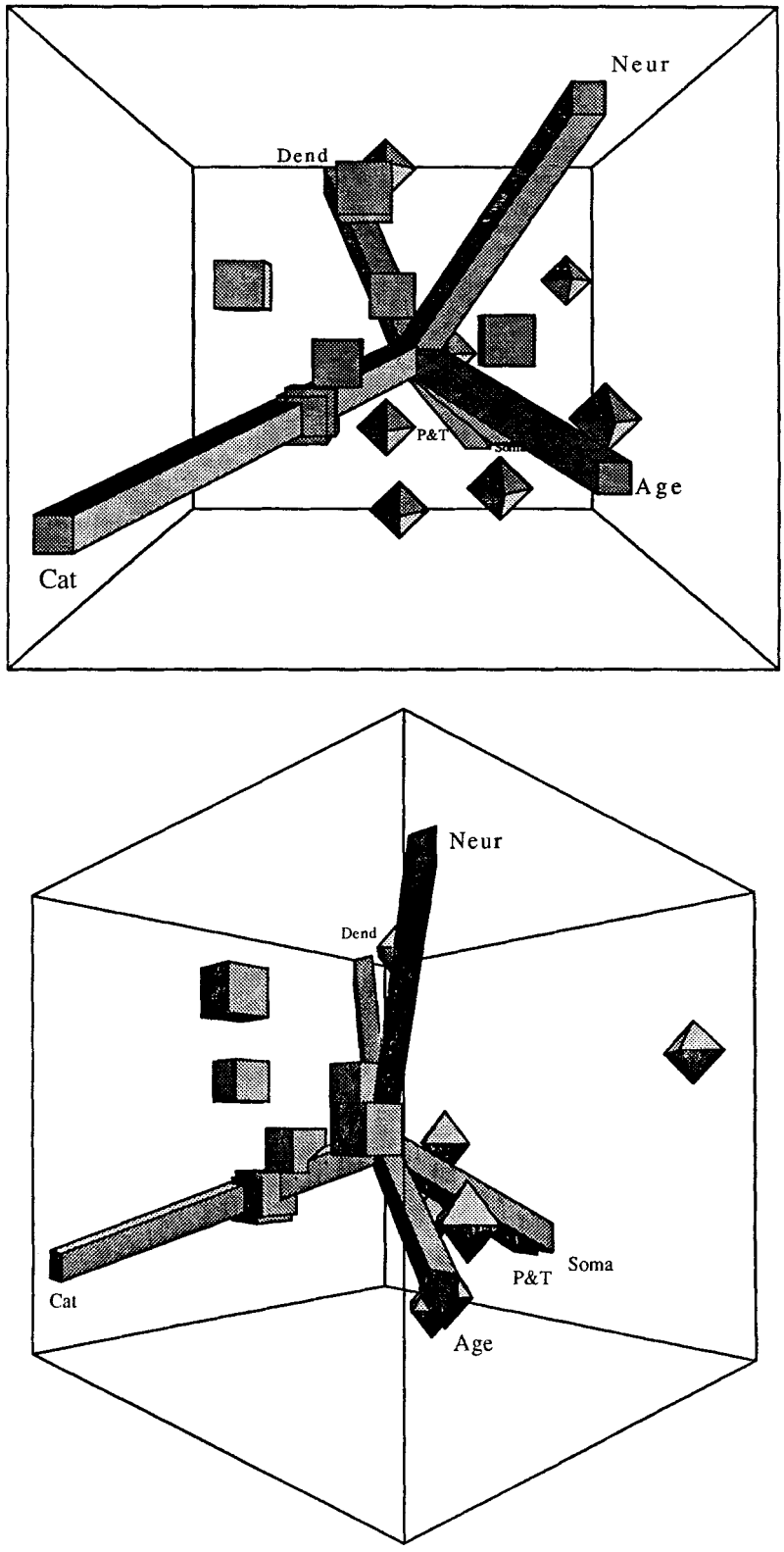

Figure 5. Measurements on Alzheimer and control brains: (a) original view (b) after 45° rotation about vertical axis. Octagons are for Alzheimer cases, cubes for controls

Table II. Coalminers classified by age and respiratory symptoms[19]

Age	Breathlessness		No breathlessness		Total
	Wheeze	No wheeze	Wheeze	No wheeze	
20–24	9	7	95	1,841	1,952
25–29	23	9	105	1,654	1,791
30–34	54	19	177	1,863	2,113
35–39	121	48	257	2,357	2,783
40–44	169	54	273	1,778	2,274
45–49	269	88	324	1,712	2,393
50–54	404	117	245	1,324	2,090
55–59	406	152	225	967	1,750
60–	372	106	132	526	1,136
Total	1827	600	1833	14,022	18,282
Overall proportion	0·0999	0·0328	0·1003	0·7670	1·000

Any investigator with moderate computational support can obtain 2-D biplots, but effective 3-D display requires special packages such as MACSPIN,[16] DATA DESK[17] or JMP[18] on the MacIntosh or a cooperative effort with a statistical or computing specialist.

A CONTINGENCY TABLE ON RESPIRATORY SYMPTOMS

Next is an example of a table of frequencies. Table II shows the number of coalminers by age (rows) and by respiratory symptoms (columns).[19] Such a contingency table differs in principle from the two-way tables – like Table I – previously biplotted. The entries of those tables were *measurements* on some variable, such as a birth rate, a pain score or a neuron count. In contingency tables, on the other hand, the entries are *counts* of how frequently the row and column categories occur together. Their statistical analysis therefore differs from that of two-way tables of measurements, and includes such techniques as a chi-square test of the independence of row and column categories. Their biplot display is also different and permits graphical approximate tests of significance, as illustrated next.

Each row of Table II represents a sample of coalminers of a certain age group, distributed into four categories according to the respiratory symptoms present. The simplest, though unlikely, model for such data would be that the symptoms were independent of age, so that one could represent their frequencies in each age group by the average frequency for all ages (shown at the bottom of Table II). The usual chi-square test for this hypothesis compares the observed frequencies with those 'expected' under the independence hypothesis. For these data the statistic is 2734, which has a minuscule P-value compared with the chi-square distribution with $(9 - 1)(4 - 1) = 24$ d.f. It is not surprising that the hypothesis of independence is not tenable.

Since symptoms were not independent of age, one needs to study the deviations of the observed frequencies from those 'expected'. Figure 6 displays these deviations in a biplot by means of markers for each age group sample and arrows for each symptom category. The biplot shows one long arrow to the left, that for the healthy category – 'no breathlessness, no wheeze'. On the right is the arrow for the worst category – 'breathlessness and wheeze'. The other two categories' arrows are in between, with 'wheeze' pointing some way up. Evidently, the horizontal direction approximately represents a healthy versus sick axis.

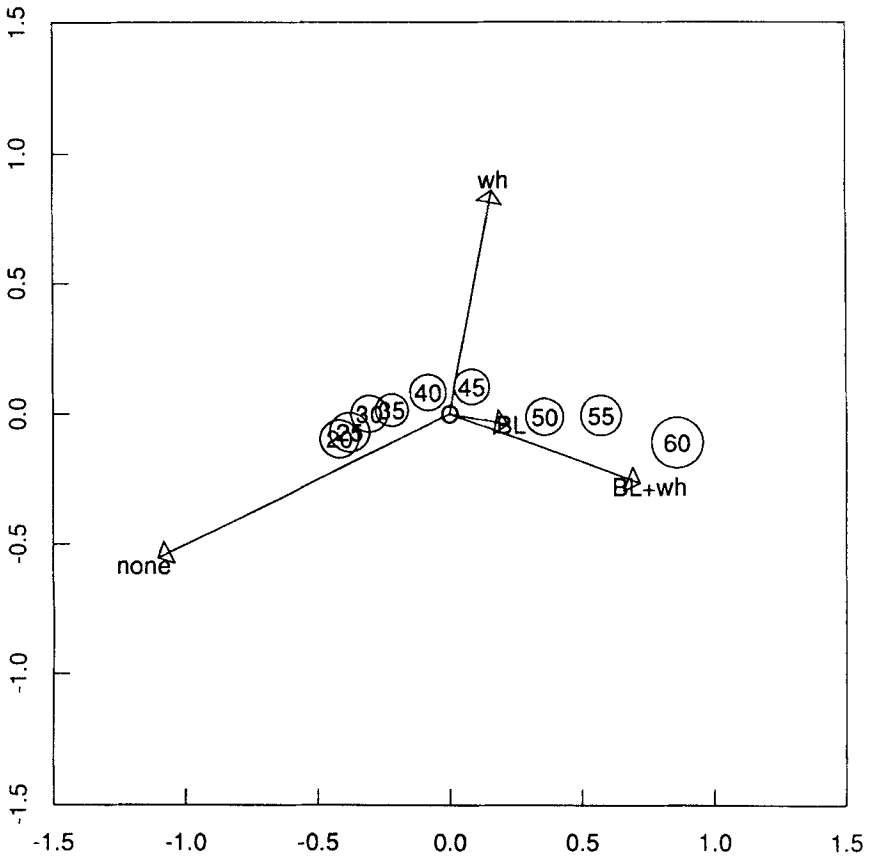

Figure 6. Coalminers: age, breathlessness and wheezing. Chi-square 2733·8, goodness of fit 99·56 per cent, uncertainty circles 99·5 per cent

The markers for the nine samples string out in order from those for the low ages on the left to those for the high ages on the right, with a slight upward curve in the intermediate ages. This is an obvious indication of a strong association of respiratory symptoms with age.

Around each age group marker is an 'uncertainty circle' that reflects the variability of that sample. These circles have been designed so that when two overlap (do not overlap), the corresponding two samples did not (did) differ significantly in their distributions into the symptom categories. Thus the two lowest age groups did not differ significantly, and nor did the next two age groups, but from age 40 onwards all the age groups differed significantly from one another.

The above circle overlap tests are approximations to the chi-square tests that compare any two samples. We ran them here at a level of 0·5 per cent experimentwise, that is we allowed 0·5 per cent chance of any false rejection among all the pairwise sample comparisons. The chance of a false rejection on a particular comparison was lower yet.

For a more detailed examination of the frequency of the different respiratory symptoms, Figure 7 is a biplot of only the first three columns of Table II, with omission of the non-symptomatic category. Again, the statistic of 536, when compared with the chi-square distribution with 18 d.f., shows a highly significant association with age.

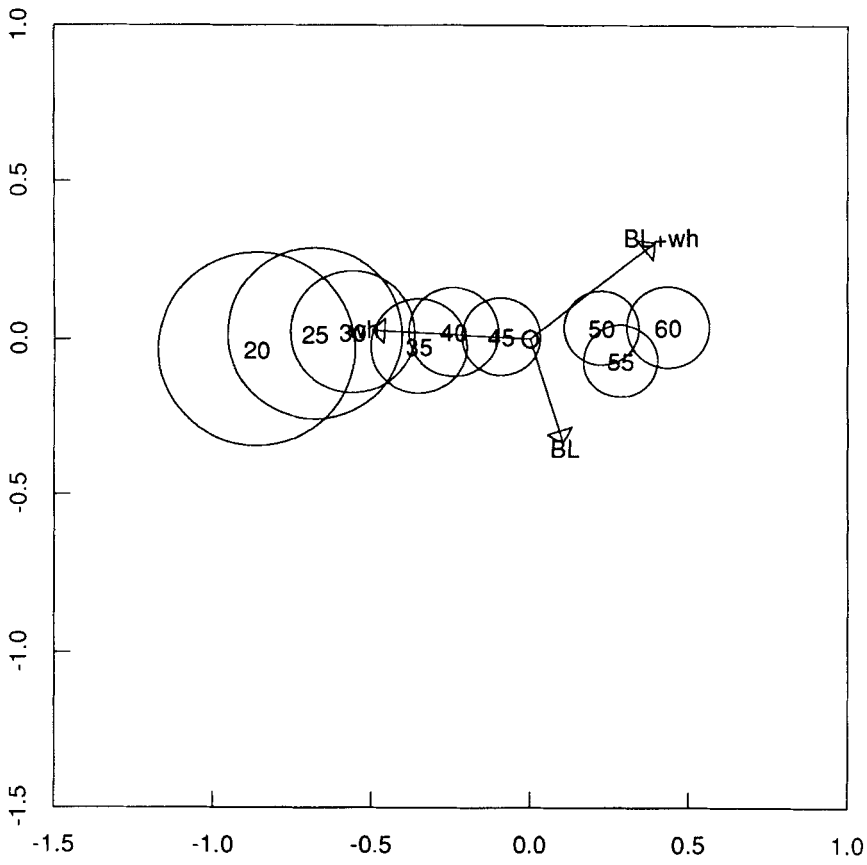

Figure 7. Symptomatic coalminers: age and symptoms. Chi-square 536·1, goodness of fit 100 per cent, uncertainty circles 99·5 per cent

The pattern in Figure 7 is one in which the horizontal direction has 'breathlessness' to the right and 'wheeze' to the left. Again, the age progression is from left to right and shows that the older were the miners, the higher was the incidence of breathlessness. 'wheezing', on the other hand, appears in a roughly vertical direction on the biplot, opposite to 'breathlessness only'. But there is little vertical variation in the age group markers, which indicates that wheezing did not vary much with age.

The uncertainty circles of Figure 7 are larger than those in Figure 6 because they are based on smaller samples (including only the symptomatic cases). Hence there is more overlap and fewer of the pairwise comparisons were significant. For example, age groups 35 and 45 differed significantly in the entire data (no overlap in Figure 6), but did not differ significantly in the symptomatic data (overlap in Figure 7).

We do not suggest substitution of these graphical checks of overlap for formal tests of significance, but they are useful in providing a quick visual indication of where one may, or may not, expect significant results.

Finally, we remark that the contingency table biplot has some affinity with correspondence analysis.[20] The latter allows comparisons of the rows with one another and of the columns with

one another, but does not permit representation of the cell frequencies or of the pairwise tests of significance.

USEFULNESS OF BIPLOTS

Biplots are useful for visualization of multivariate observations and assessment of the features of data. They can reveal patterns of correlation and covariation, and of differences between groups of sample units. Most importantly, they show the relation between individual units and the multivariate structure, and indicate which units differ (or are similar) on what variables.

Biplots and other graphical displays should supplement, rather than supplant, formal statistical analyses. Formal analyses focus on explicit hypotheses and use specific models; graphics allow more tentative exploration and require few assumptions. If one knew all about the underlying probability model and sought definitive confirmation or rejection of hypotheses, one would not need to go beyond classical statistical inference methods. But when one is hesitant about the use of restrictive assumptions, such as normality or equality of variances, one should begin by inspecting the data graphically. And again, after finding a model, such as a linear experimental design or a regression, it is often advisable to plot the residuals, both to check whether one might need to change the model and to see whether additional features emerge. In both stages of study – in the initial overview of the data and in the final inspection of residuals – one may find biplots particularly helpful. There are numerous examples of their uses in biomedical research.[2,9,10,14,21–27]

Biplots are simple in concept and in interpretation. They use the viewer's native tendencies to refer to axes for coordinates and to associate nearby points and closely angled lines as showing similarity. Their construction entails a certain amount of mathematics (see appendix), but their use to gain a better understanding of data requires little sophistication.

APPENDIX: SOME MATHEMATICS AND SOFTWARE

The CMP biplot[1–3]

Let Y denote the $n \times m$ matrix to be displayed, its columns already centred and scaled appropriately. The CMP biplot then displays row markers

$$\mathbf{a}_i = (a_{i,1}, a_{i,2}), \qquad i = 1, \ldots, n$$

and column markers

$$\mathbf{b}_j = (b_{j,1}, b_{j,2}), \qquad j = 1, \ldots, m.$$

These result from the singular value decomposition

$$Y = P' \Lambda Q$$

by defining A and B as the matrices of the first two columns of P' and of $Q'\Lambda$, respectively. Then $\mathbf{a}_i'$ is the ith row of A and $\mathbf{b}_j'$ is the jth row of B.

The above singular value decomposition could be obtained by computing the decomposition

$$Y'Y = Q'\Lambda^2 Q$$

into eigenvalues (elements of diagonal Λ^2) and eigenvectors (row of Q) and then using

$$P' = YQ'\Lambda^{-1}$$

The contingency table biplot

Consider n samples classified into m categories, and write $f_{i,j}$ for the frequency of sample i in category j. Then the size of the ith sample is

$$f_{i*} = \Sigma_j f_{i,j}$$

and the overall proportion in the jth category is

$$p_{.j} = \Sigma_i f_{i,j} / \Sigma_i \Sigma_j f_{i,j}.$$

As usual in contingency table analysis, define the frequency 'expected' in cell (i, j) as

$$e_{i,j} = f_{i*} p_{.j}$$

and the standardized residual as

$$y_{i,j} = (f_{i,j} - e_{i,j}) / \sqrt{e_{i,j}}.$$

With this notation, the usual chi-square statistic becomes $\Sigma_i \Sigma_j y_{i,j}^2$.

The biplot coordinates obtain from the singular value decomposition of Y by setting

$\mathbf{a}_i = \sqrt{(1/f_{i*})}$ times the first two elements of the ith row of $P'\Lambda$

$\mathbf{b}_j = \sqrt{(p_{.j})}$ times the first two elements of the jth row of Q'.

The uncertainty circle for the ith sample centres on row marker $\mathbf{a}_i$ and has radius d_i, where

$$d_i^2 = \chi^2_{(m-1)\alpha} \{1 + k_i\} / \{(1 + \sqrt{k_i})^2 f_{i*}\}$$

with

$$k_i = \max \{f_{i*} / \min_e f_{e*}, \max_e f_{e*} / f_{i*}\}$$

$$\alpha = (\text{experimentwise level}) / \{m(m-1)/2\}.$$

One can use the above formulae to compute the coordinates of row markers and column arrowheads. Alternatively, one can use any PCA routine, with appropriate scaling, to obtain these coordinates. After choice of suitable scales, one can use the coordinates to draw the biplot.

In many data sets some entries are missing, and before biplotting one needs to interpolate for missing values. A possible method is to obtain multiple regression equations for the set of complete observations and to use them to 'predict' the missing observation on a unit from the other observations that are available on that unit. A more direct method was proposed by Gabriel and Zamir.[12]

Although one could plot by hand, one would be more likely to produce the plot with some computer graphics device which may also allow special effects such as colour labels[6,7,17,18] or real time rotation of 3-D biplots.[7,16-18]

A number of special routines are available that produce biplots directly from the data matrices.[4] In particular, these include macros for MINITAB which have wide applicability but produce less than perfect graphic output. Excellent graphics with great flexibility of labelling and display can be obtained by S macros. Some of the other biplot possibilities are in using JMP, GENSTAT or DATA DESK. The user should have little trouble in adapting any flexible package to the display of biplots.

ACKNOWLEDGEMENTS

An early version of this work was presented at the Biometrics Society (ENAR/WNAR) meeting in March 1987. The present version was revised by the first author after Charles Odoroff's death on 29 December 1987, and edited by Setta Odoroff. The work was supported, in part, by the Office of Naval Research under contract N00014-80-C-0387 on multivariate graphics, and by the National Cancer Institute through grant NCI 5P30 CA11198-20.

REFERENCES

1. Gabriel, K. R. 'The biplot graphical display of matrices with applications to principal component analysis', *Biometrika*, **58**, 453–467 (1971).
2. Gabriel, K. R. 'Biplot display of multivariate matrices for inspection of data and diagnosis', in Barnett, V. (ed.), *Interpreting Multivariate Data*, Wiley, London, 1981, pp. 147–173.
3. Gabriel, K. R. 'Biplot', in Kotz, S., Johnson, N. L. and Read, C. (eds), *Encyclopedia of Statistical Sciences*, volume I, Wiley, New York, 1981, pp. 262–265.
4. Gabriel, K. R. and Odoroff, C. L. 'The biplot for exploration and diagnosis: examples and software', University of Rochester, Statistics and Biostatistics Technical Report 86/03, 1986.
5. Gabriel, K. R., Rave, G. and Weber, E. 'Graphische Darstellung von Matrizen durch das Biplot', *EDV in Medizin und Biologie*, **7**, 1–15 (1976).
6. Gabriel, K. R. and Odoroff, C. L. 'Use of 3D biplots for diagnosing models to fit higher dimensional data', in Wegman, E. and DePriest, D. (eds), *Statistical Image Processing and Graphics*, Dekker, New York, 1986, pp. 257–274.
7. Odoroff, C. L., Basu, A., Gabriel, K. R. and Therneau, T. M. 'ANIMATE: an interactive color statistical graphics system for three dimensional displays', National Computer Graphics Association, *Proceedings of the Seventh Annual Conference*, volume III, 1986, pp. 723–731.
8. Baaske, W. *Hauptebenenanalyse*, Studiengruppe fuer Internationale Analysen, Laxenburg, Austria, 1985.
9. Cox, C., Davis, H. T., Wardell, W. M., Calimlim, J. F. and Lasagna, L. 'Graphical analysis of multivariate pain data in analgesic trials', *Controlled Clinical Trials*, **7**, 53–64 (1986).
10. Cox, C. and Gabriel, K. R. 'Some comparisons of biplot display and pencil and paper EDA methods', in Launer, R. L. and Siegel, A. F. (eds), *Modern Data Analysis*, Academic Press, New York, 1982, pp. 45–82.
11. Gabriel, K. R. and Odoroff, C. L. 'Resistant lower rank approximation of matrices', in Diday, E., Jambu, M., Lebart, L., Pages, J. and Tomassone, R. (eds), *Data Analysis and Information 3*, North-Holland, Amsterdam, 1984, pp. 23–30.
12. Gabriel, K. R. and Zamir, S. 'Lower rank approximation of matrices by least squares with any choice of weights', *Technometrics*, **21**, 489–449 (1979).
13. Gabriel, K. R. and Odoroff, C. L. 'Some reflections on strategies of modelling: how, when and whether to use principal components', in Sen, P. K. (ed.), *Biostatistics – Statistics in Biomedical, Public Health and Environmental Science*, North-Holland, Amsterdam, 1985, pp. 315–331.
14. Strauss, J. S., Gabriel, K. R., Kokes, R. F., Ritzler, B. A., VanOrd, A. and Tarana, E. 'Do psychiatric patients fit their diagnoses? Patterns of symptomatology as described with the biplot', *Journal of Nervous and Mental Disease*, **167**, 105–113 (1979).
15. Hamill, R. W., Caine, E. D., Coleman, P. D., Eskin, T. A., Flood, D. G., Joynt, R. J., Lapham, L. W., McNeil, T. H. and Odoroff, C. L. 'Multiple morphological and biochemical measures completely distinguish patients with Alzheimer's disease', *Society for Neuroscience Abstracts*, **13**, 442 (1987).
16. Donoho, A., Donoho, D. and Gasko, M. *MACSPIN Graphical Data Analysis Software*, D^2 Software Inc., Austin, Texas, 1986.
17. Velleman, P. F. and Velleman, A. Y. *Data Desk Handbook*, Odesta Corporation, Illinois, 1988.
18. SAS Institute, *JMP User's Guide*, SAS Institute, Cary, NC, 1989.
19. Ashford, J. R. and Sowden, R. R. 'Multi-variate probit analysis', *Biometrics*, **26**, 535–546 (1970).
20. Greenacre, M. J. *Theory and Applications of Correspondence Analysis*, Academic Press, London, 1984.
21. Berg, R. L., Brooks, M. R. and Savicevic, M. *Health Care in Yugoslavia and the United States*, DHEW (NIH), Washington DC, 1976.
22. Bradu, D. and Grine, F. E. 'Multivariate analysis of diamodontine crania from South Africa and Zambia', *South African Journal of Science*, **75**, 441–448 (1979).

23. Kempton, R. A. 'The use of biplots in interpreting variety by environment interactions', *Journal of Agricultural Science (Cambridge)*, **103**, 123–135 (1984).
24. Kunitz, S. J. *Disease Change and the Role of Medicine*, University of California Press, Berkeley, 1983.
25. Ohmayer, G. and Seiler, H. 'Numerische Gruppierung und graphische Darstellung von Daten: Ein Methodenvergleich', *EDV in Medizin und Biologie*, **16**, 65–73 (1985).
26. Osmond, C. 'Biplot models applied to cancer mortality rates', *Applied Statistics*, **34**, 63–70 (1985).
27. Shy-Modjeska, J. S., Riviere, J. and Rawlings, J. O. 'Application of biplot methods to the multivariate analysis of toxicological and pharmacokinetic data', *Toxicology and Applied Pharmacology*, **72**, 91–101 (1984).