AD-A114 170 ROCHESTER UNIV NY F/G 12/1

SOME COMPARISONS OF BIPLOT DISPLAY AND PENCIL-AND-PAPER E.D.A. --ETC(U)

SEP 81 C COX, K R GABRIEL N00014-80-C-0387

UNCLASSIFIED TR-81/22 NL

1 OF 1
AD
A114170

END
DATE
FILMED
5-82
DTIC

SOME COMPARISONS OF BIPLOT DISPLAY AND PENCIL-AND-PAPER E.D.A. METHODS

Revision

by

Christopher Cox

and

K. Ruben Gabriel

TECHNICAL REPORT 81/22

June 1980
Revised September 1981

Department of Statistics and Division of Biostatistics
University of Rochester, Rochester, New York 14642, USA

Research supported in part by ONR contract N 0014-80-C-0387 on Biplot Multivariate Graphics.

72

SOME COMPARISONS OF BIPLOT DISPLAY AND PENCIL-AND-PAPER E.D.A. METHODS[1]

Christopher Cox
K. Ruben Gabriel

Department of Statistics and
Division of Biostatistics
University of Rochester
Rochester, New York

Presented at ARO Workshop on Modern-Data Analysis at
Raleigh, North Carolina
June 2-4, 1980

This paper presents some comparisons of EDA and biplot display. By pencil-and-paper EDA we mean the methods advocated by John Tukey in his 1977 volume (Tukey, 1977). We use examples from that book to illustrate the differences and the similarities of the two methods. We assume that the analyses in the book are familiar and show how our biplot analyses differ from them.

The paper begins with an introduction to the biplot, accompanied by one example, in which the biplot is used for data summarization and description. Then we look at four diagnostic examples from the book and show what biplot display would have done. We end by drawing some conclusions. References for further reading on the biplot and its diagnostic and

[1] Research supported in part by ONR contract N 00014-80-C-0387 on Biplot Multivariate Graphics.

other uses are included at the end of the paper. Computer programs are available from the authors at the Division of Biostatistics of the University of Rochester.

We start by explaining what the biplot is. It is a graphical display of a matrix $Y_{(nxm)}$ of n rows and m columns by means of row markers $\underline{a}_1, \underline{a}_2, \ldots \underline{a}_n$ and column markers $\underline{b}_1, \underline{b}_2, \ldots \underline{b}_m$. The biplot carries one marker for each row, and one marker for each column. The principle of biplot display of a matrix Y is that element $y_{i,j}$ in the i-th row and j-th column is represented by the inner product of the i-th row marker and the j-th column marker, i.e., $\underline{a}_i'\underline{b}_j$ represents $y_{i,j}$. A 100 by 20 matrix, for example, would be represented by 100 row markers and 20 column markers in such a way that all 2,000 elements are represented by inner products of row markers and column markers. To set this in matrix terms we may array the row markers $\underline{a}_i$ as rows of matrix A and the column markers $\underline{b}_j$ as columns of a matrix B'. Clearly, then the matrix product AB' represents the matrix Y itself.

On a point of terminology, the prefix "bi-" of biplot serves to indicate that this is a joint display of rows and columns. It does not indicate the two-dimensionality of the biplot. Any plot is two dimensional. On the other hand, if we use a three-dimensional display analogous to the biplot, we call it a bi-model because it too is a joint display of both rows and columns: the ending "model" indicates that there are three dimensions.

Figure 1 shows a very simple example of a biplot. Y is a 4 x 3 matrix of rank 2; the row markers are the rows of matrix A, and the column markers are the columns of matrix B'. Each row of A and each column of B' is displayed on this biplot--

Legend: ● $\underline{a}_u$ is u-th row marker; ↗ $\underline{b}_v$ is v-th column marker

$$\begin{array}{ccccc} Y & = & A & & B' \\ \begin{bmatrix} 2 & 2 & -4 \\ 2 & 1 & -3 \\ 0 & -1\frac{1}{2} & 1\frac{1}{2} \\ -1 & -\frac{1}{2} & 1\frac{1}{2} \end{bmatrix} & = & \begin{bmatrix} 2 & 2 \\ 2 & 1 \\ 0 & -1\frac{1}{2} \\ -1 & -\frac{1}{2} \end{bmatrix} & & \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix} \end{array}$$

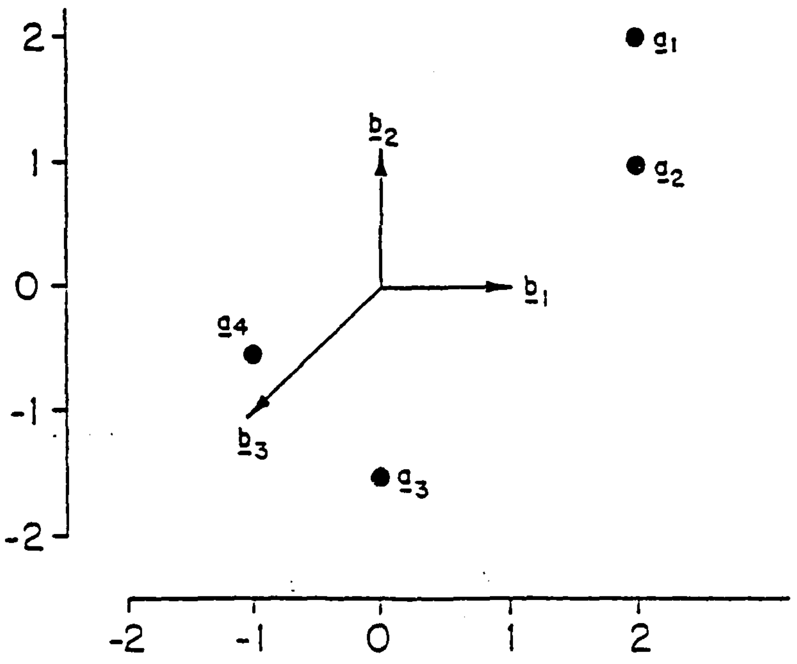


Fig. 1. A matrix Y, its factorization AB' and the biplot.

seven markers in all. For convenience the row markers $\underline{a}_i$ are indicated as circles whereas the column markers $\underline{b}_j$ are indicated as vectors.

Figure 2 illustrates how particular elements of the matrix are represented on the biplot. Thus, element $y_{2,3}$ is represented by the inner product of the second row marker $\underline{a}_2$ and the third column marker $\underline{b}_3$. To see this geometrically, we choose one of these markers, e.g., $\underline{b}_3$, take the straight line

$$\left\{ \begin{array}{l} y_{2,3} = -(\text{Length of } \underline{b}_3) \times (\text{Length of projection of } \underline{a}_2 \text{ onto } \underline{b}_3) \\ y_{4,3} = (\text{Length of } \underline{b}_3) \times (\text{Length of projection of } \underline{a}_4 \text{ onto } \underline{b}_3) \end{array} \right\}$$

$$\begin{array}{cccc} \text{Third column of Y} \quad \underline{y}_{(3)} & = & A & \underline{b}_3 \\ \begin{bmatrix} -4 \\ -3 \\ 1\frac{1}{2} \\ 1\frac{1}{2} \end{bmatrix} & = & \begin{bmatrix} 2 & 2 \\ 2 & 1 \\ 0 & -1\frac{1}{2} \\ -1 & -\frac{1}{2} \end{bmatrix} & \begin{bmatrix} -1 \\ -1 \end{bmatrix} \end{array}$$

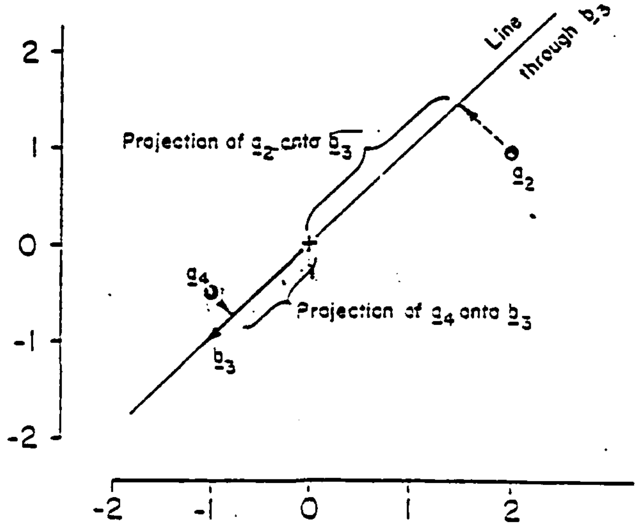

Line through $\underline{b}_3$
Projection of $\underline{a}_2$ onto $\underline{b}_3$
$\underline{a}_2$
$\underline{a}_4$
Projection of $\underline{a}_4$ onto $\underline{b}_3$
$\underline{b}_3$
2
1
0
-1
-2
-2
-1
0
1
2

Fig. 2 Biplot representation of the third column of Y.

from the origin through that marker, and project the other marker $\underline{a}_2$ orthogonally onto it. The distance from the origin to the foot of the perpendicular of $\underline{a}_2$ onto the line through $\underline{b}_3$ is then multiplied by the length of the vector $\underline{b}_3$ to obtain the inner product. For another example, take $y_{3,3}$ for which we project $\underline{a}_3$ onto $\underline{b}_3$. In this case, the projection is half as long as the one before and in the opposite direction, i.e., in the direction of $\underline{b}_3$ itself. So the product is positive and half the size of the previous one.

A few more remarks about biplots need to be made. First of all note that the biplot is planar--the row markers $\underline{a}_i$, as well as the column markers $\underline{b}_j$, are plotted in the plane. This cannot be done exactly for any matrix of rank greater than 2. Hence, the first step for biplotting such a matrix Y is to approximate it by a matrix $Y_{[2]}$ of rank 2. This is called lower rank approximation. The second step is to factorize the $Y_{[2]}$ approximation into a product AB' of an A matrix of two columns and a B' matrix of two rows. Then the rows of A can be plotted as row markers $\underline{a}_i$ and the columns of B' as column markers $\underline{b}_j$. Their joint display is a biplot. Biplotting is thus seen to require three steps: rank 2 approximation, factorization, and display.

Lower rank approximation can be carried out by means of the theorem due to Householder and Young (1938), which provides the least squares solution to this problem. When weights are introduced, and each of the squared differences $(y_{ij} - \underline{a}_i'\underline{b})^2$ is to be weighted by some given w_{ij}, the mathematics of that theorem break down. However, a weighted least squares algorithm and suitable initialization methods are available (Gabriel and Zamir, 1979). An earlier solution for particular kinds of weights common in statistics was provided by Haber (1975). Another method of reduced rank approximation uses adaptive fits (McNeil and Tukey, 1975). C. L. Odoroff of Rochester is currently working on using the weighted least squares solution for adaptive fitting, i.e., taking the residuals from the last fit and using them to adjust the weights for the next fit.

We now turn to uses of the biplot. These are mostly of two kinds--inspection of data and diagnostics. We present one brief example of biplot inspection before we go on to our main subject which is biplot diagnostics.

We consider data from single-dose, postoperative, oral analgesic trials. Patients who had previously consented to participate, and who requested medication for moderate to severe pain during the first three days after surgery, were given a single dose of one of the study drugs on a randomized, double-blind basis. The resulting data consist of ordinal pain scores, on a five point scale, with zero being no pain and four being very severe pain. (A number of standard pain scales were used, of which we have chosen this one as an illustration.) Data were recorded at baseline (medication time), one-half hour later, and at hourly intervals until five hours after medication--Figure 3. There were a total of 180 patients in the trials, with eight treatments, including a placebo.

In Figure 4 we show a portion of a biplot of the data matrix. We have included only two of the treatment groups: those receiving a placebo and those given a highly effective

Surgical Patients	Baseline (Medication Time)	Time after medication					
		1/2 hr.	1 hr.	2 hrs.	3 hrs.	4 hrs.	5 hrs.
1	. . . .	. .	. .	. .	. .	. .	. .
2	. . . .	. .	. .	. .	. .	. .	. .
.							
.				DATA			
180	. . . .	. .	. .	. .	. .	. .	. .

Note: Ordinal pain scores on a 5-point scale from 0 (no pain) to 4 (worst pain I have ever experienced).

Fig. 3. Data from single dose postoperative oral analgesic trials.

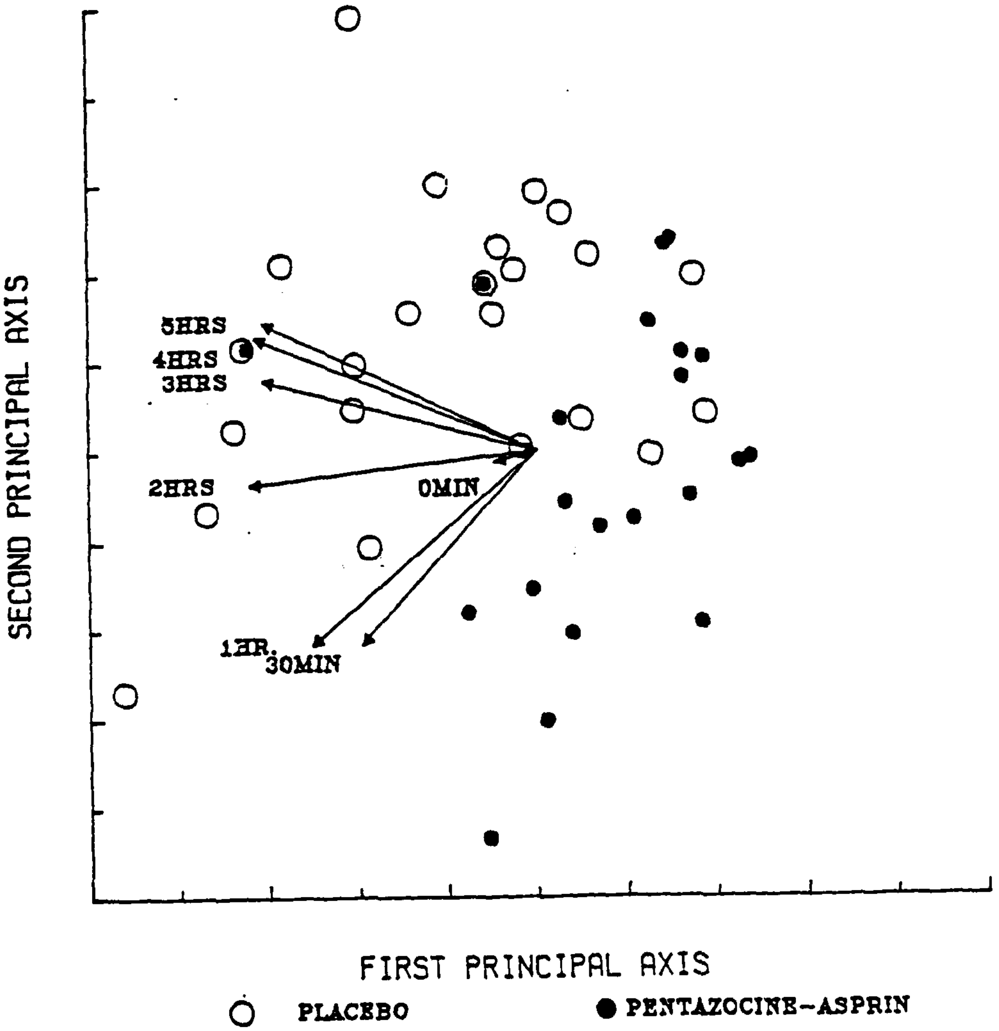


Fig. 4. Biplot of ordinal pain scores from two treatment groups in analgesic trials.

combination analgesic (pentazocine and aspirin). This particular biplot has been scaled so that the lengths of the arrows represent standard deviations and the angles between the arrows represent correlations among the corresponding columns--times of recording. (This is *referred* to as a *GH'* biplot--Gabriel, 1971.)

Figure 5 replaces the row markers of Figure 4 by one standard deviation concentration ellipses. Each ellipse summarizes the row markers of the approximately 25 patients in the corresponding treatment group. The center of each ellipse is also plotted.

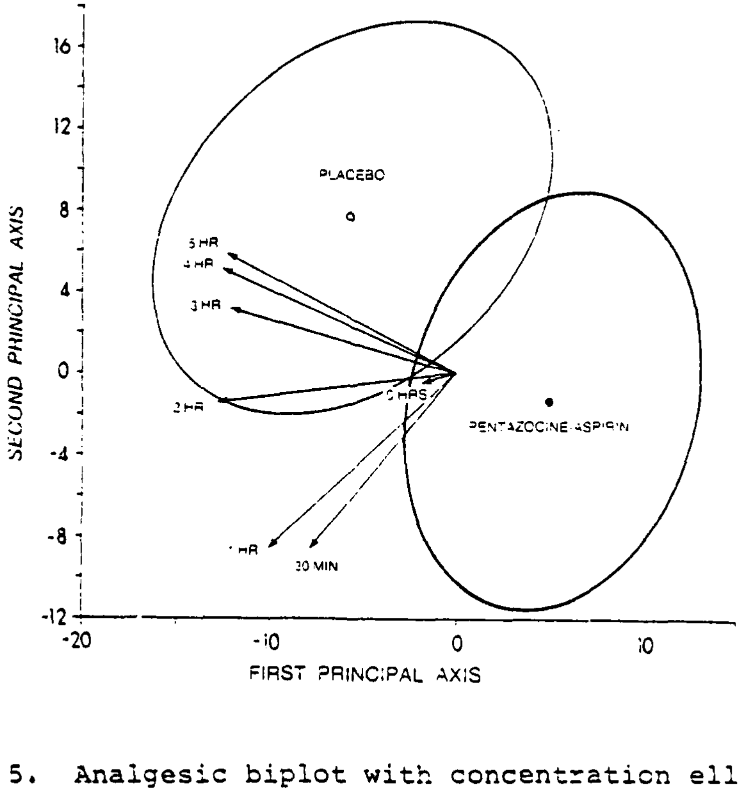
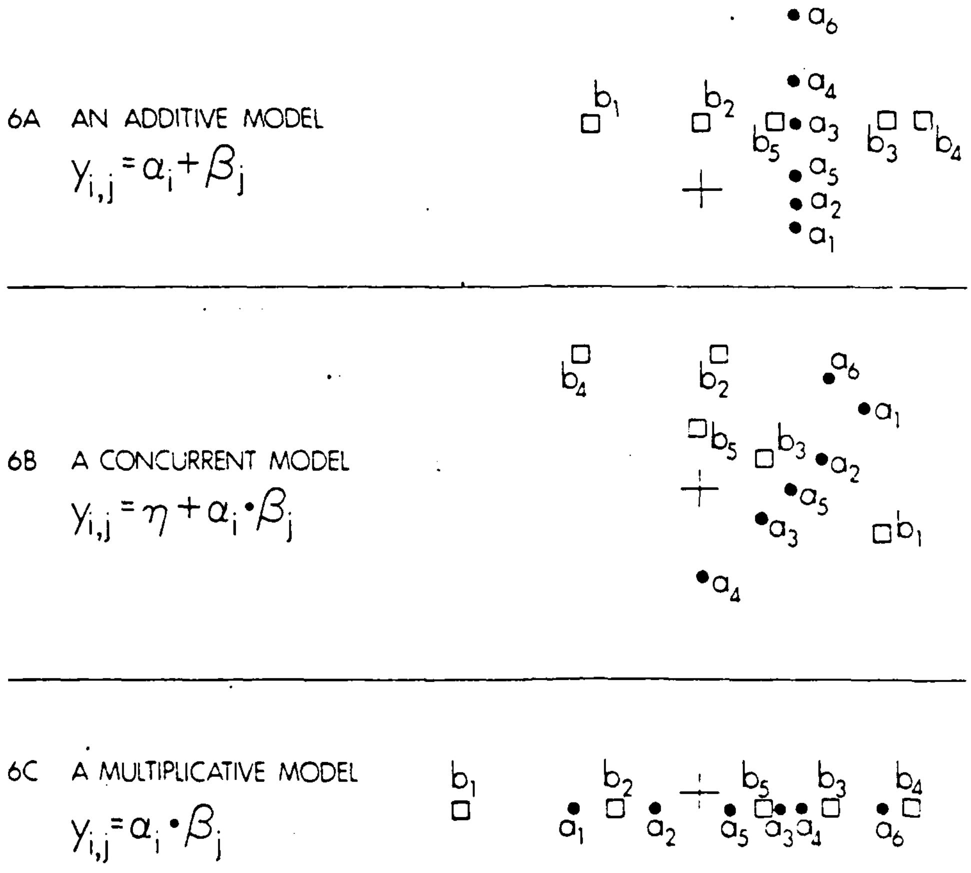


Fig. 5. Analgesic biplot with concentration ellipses summarizing treatment groups.

The small angle between the arrows for the one-half hour and one hour pain scores shows that scores at those times are fairly well correlated with one another. Similarly, one can see that the three, four, and five hour pain scores are correlated with each other, but roughly uncorrelated--arrows at about 90°--with the scores at earlier times. The shortness of the baseline arrow reflects the fact that only patients with baseline pain scores of 2 or 3 were included in these trials. (Further examination of the data suggests that the baseline pain scores are not well represented by the biplot.) In order to see the effects of the two treatments, we examine the average pain scores of the two treatment groups at different time points by projecting the centers of the ellipses onto the arrows. We see that both groups were similar (and below the overall average) at the one-half and 1 hour time points; this is an indication of a placebo effect of surprising duration. With respect to the later time points, however, the placebo group had much higher than average pain scores, while the pentazocine-aspirin group had below average pain scores. (The other six treatment groups were intermediate, increasing in efficacy from placebo.) Clearly, the later time points are more sensitive to the effect of analgesics. This analysis contrasts with the traditional method of analyzing such data in which the post-medication pain scores are cummulated to obtain a measure of "total pain" and the time differentials are ignored (Cox et al, 1980). Inspection of this biplot has suggested new derived pain measures, which appear to be more sensitive to treatment differences.

We now turn to the second use of the biplot which is to facilitate the search for patterns and the inference of models to fit the data. To illustrate, some of the patterns that we look for are shown in Figure 6. Figure 6A shows row markers and column markers which are both collinear and have a right angle between their two lines. Such a biplot pattern indicates that the data are well fitted by an additive model, i.e., $y_{i,j} = \alpha_i + \beta_j$ for some row effects α_i and some column effects β_j. This is something that the eye picks up readily: row markers on a line, column markers on a line, and a 90° angle between them.



6A AN ADDITIVE MODEL $y_{i,j}=\alpha_i+\beta_j$

6B A CONCURRENT MODEL $y_{i,j}=\eta+\alpha_i \cdot \beta_j$

6C A MULTIPLICATIVE MODEL $y_{i,j}=\alpha_i \cdot \beta_j$

Fig. 6. Biplot patterns and the models which they diagnose.

Another biplot pattern, shown in Figure 6B, has the column markers aligned, and the row markers aligned, but the angle between the lines is not 90°. A concurrent model can then be diagnosed. This is also known as a degree of freedom for non-additivity model and it can be parametized most simply as $y_{i,j} = \eta + \alpha_i \beta_j$.

Finally, Figure 6C has all markers on one line. This obviously diagnoses a model of rank one, i.e., $y_{i,j} = \alpha_i \beta_j$.

It will have been noticed that in Figure 6B there is one column marker--$\underline{b}_3$--that is not aligned with the other $\underline{b}$'s and one row marker--$\underline{a}_4$--that is not aligned with the other $\underline{a}$'s. That indicates that the third column and fourth row are not fitted by the concurrent model, though the other rows and columns are. This illustrates a very useful property of the biplot; if a pattern fits only some of the column markers and some of the row markers, the implied model may be diagnosed exclusively for those columns and rows. Indeed, a biplot is not only a display of the whole matrix, but can also be regarded as a simultaneous display of all possible submatrices. The eye immediately picks up subsets and subtables and allows their separate diagnosis.

We should add that outlying rows or columns might at times distort the rank two approximation and spoil the chances of diagnosing a model. There might also be situations where the subtable models cannot be seen on the biplot because the biplot mainly displays subtable differences. In such cases it might be helpful to employ a 3D bimodel and see whether any simple patterns are evident on some projection of such a bimodel.

Row Markers $\underline{a}_i$	Column Markers $\underline{b}_j$	The Model for $y_{i,j}$ is:
Collinear	-	$\beta_j + \alpha_i\delta_j$ columns regression
-	Collinear	$\alpha_i + \gamma_i\beta_j$ rows regression
Collinear	Collinear	$\mu + \gamma_i\beta_j$ one degree of freedom for non-additivity
Collinear lines 90° to each other	Collinear	$\alpha_i + \beta_j$ additive

Fig. 7. Some biplot diagnostic rules (Bradu and Gabriel, 1978).

The examples of Figure 6 illustrate some simple diagnostic rules which are listed more formally in Figure 7. There are four collinearity patterns for row and/or column markers and Figure 7 shows the model that may be diagnosed from each one. Thus, collinear row markers indicate that each column can be modelled by a linear regression on some "row effects" α_i. An analogous diagnosis follows from column marker collinearity.

TABLE I. Monthly Mean Temperatures (°s F)[a]

	Caribou	Washington, D.C.	Laredo
Jan.	8.7	36.2	57.6
Feb.	9.8	37.1	61.9
Mar.	21.7	45.3	68.4
Apr.	34.7	54.4	75.9
May	48.5	64.7	81.2
Jun.	58.4	73.4	85.8
Jul.	64.0	77.3	87.7

[a] From J. W. Tukey, EDA, Chapter 10.

Joint collinearity diagnoses concurrence or additivity, depending on the angle between the lines, as already illustrated in Figure 6.

We now turn to the first example from John Tukey's book. Table I shows monthly mean temperatures at three locations, one up North, one mid-way and one further South. The biplot of the data is shown in Figure 8. The biplot column markers for the three cities are clearly collinear and the row markers for months are also pretty close to collinear. The angle between the lines is not 90°, and this suggests a concurrent

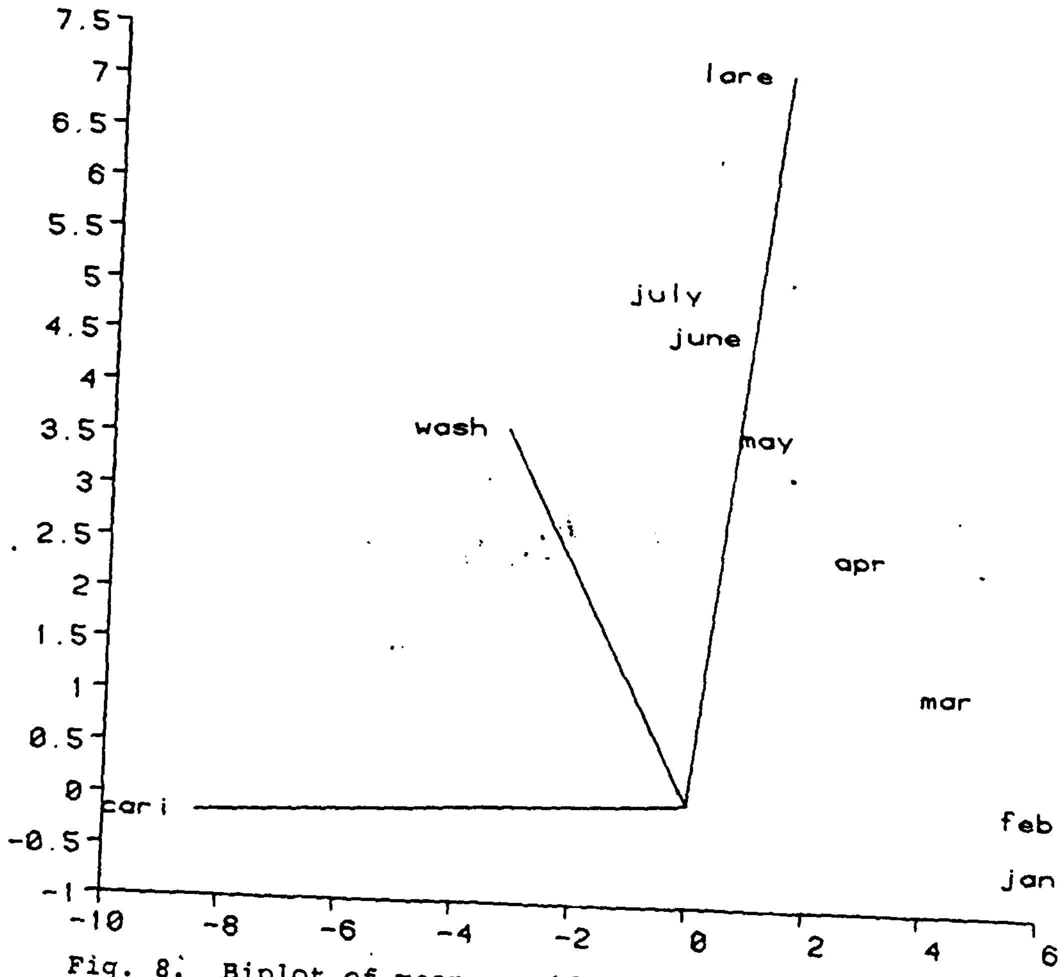


Fig. 8. Biplot of mean monthly temperature data from three cities (Caribou; Washington, D.C.; Laredo).

model. Indeed, that is exactly what Tukey concluded in his book, where he calls it a "plus-one-fit".;

It is evident that the biplot has revealed this model very simply and strikingly. Actually, a few more things may be said about this example. The months really are not quite collinear--they seem to curve around in a sequence from January to July. This leads one to wonder about addition of the remaining months. Tsianco (1980) has done similar biplots on data of 50 weather stations for twenty-four successive months. When one looks at part of his biplots, they look much like Figure 8, but when one looks at a bimodel of the entire years' data, the month markers are found to be on an ellipse in 3D. It can be shown that an ellipse on the biplot diagnoses a harmonic model for the data. That is a more reasonable model for temperature than a "concurrent", or plus-one-fit, model.

This example is considered again in Chapter 9 of Mosteller and Tukey (1977), where a one parameter family of "matched" exponential transformations is used essentially to obtain additivity (e.g., d = 70 in Exhibit 21). The biplot of the transformed data, however, still suggests a concurrent model, even though the median polish residuals do not suggest this. Thus, the biplot can serve as a useful check whether a transformation has achieved its purpose.

As a second example of the use of these diagnostic rules, we consider data on the world's supply of telephones -- Table II A--analyzed in Tukey's (1977), Chapter 12. The world was divided into seven "continents", and yearly counts were given from 1951 to 1961, with the years 1952 through 1955 omitted. Yearly increases are seen to be more proportional

TABLE II A. World's Telephones (Raw Counts).[a]

	1951	1956	1957	1958	1959	1960	1961
N.Amer.	45939	60423	64721	68484	71799	76036	79831
Eur.	21574	29990	32510	35218	37598	40341	43173
Asia	2876	4708	5230	6062	6856	8220	9053
S.Amer.	1815	2568	2695	2845	3000	3145	3338
Oceania	1646	2366	2526	2691	2868	3054	3224
Africa	895	1411	1546	1663	1769	1905	2005
MidAmer.	555	733	773	836	911	1008	1076

[a] From J. W. Tukey, EDA, Chapter 12; originally The World's Telephones 1961, American Telegraph and Telephone Company.

TABLE II B. Log$_e$ Counts of World's Telephones

	1951	1956	1957	1958	1959	1960	1961
N.Amer.	10.735	11.009	11.078	11.134	11.182	11.239	11.288
Eur.	9.979	10.309	10.389	10.469	10.535	10.605	10.673
Asia	7.964	8.457	8.562	8.710	8.833	9.014	9.111
S.Amer.	7.504	7.851	7.899	7.953	8.006	8.054	8.113
Oceania	7.406	7.769	7.834	7.898	7.961	8.024	8.078
Africa	6.797	7.252	7.343	7.416	7.478	7.552	7.603
MidAmer.	6.319	6.597	6.650	6.729	6.815	6.916	6.981

than additive, and we follow Tukey's suggestions and consider the logarithms--Table II B.

We first examine a biplot of the mean-centered log counts, shown in Figure 9. In addition to plotting the row and column markers, we have also included their arithmetic averages gmn and hmn, for row and column markers, respectively. From the evident collinearity of the column markers we diagnose a rows regression model--second row of Figure 7. The linearity of regression on time is checked by comparing the distances between column markers with the corresponding time intervals. It is thus evident from the figure that the regression is linear in time.

We next show how to use the biplot to obtain approximate parameter estimates, and thus more specific diagnoses. (For details see Bradu and Gabriel, 1978.) We first draw the line through the column markers and project the row markers orthogonally onto it. The distances from these projections to the projection of the mean of the row markers (gmn) are proportional to the estimates of the regression coefficients γ_i. (The projection of the mean gives the positive direction.) On the basis of these projections, we decide to fit a single slope for North, South, and Mid-America, and Oceania. We also require the same slope for Europe as Africa. A higher slope is clearly needed for Asia.

Further diagnosis can be obtained by projecting the row markers onto the line through the origin and the mean of the column markers (hmn). Distances from the projection of the mean (gmn) approximate the row effects (α_i) proportionally. These are the intercepts of the regressions. Indeed, the ordering of these projections is quite similar to that of the log counts in 1951. Thus, for example, North America had more telephones than Europe in 1951, but subsequently Europeans acquired them more rapidly.

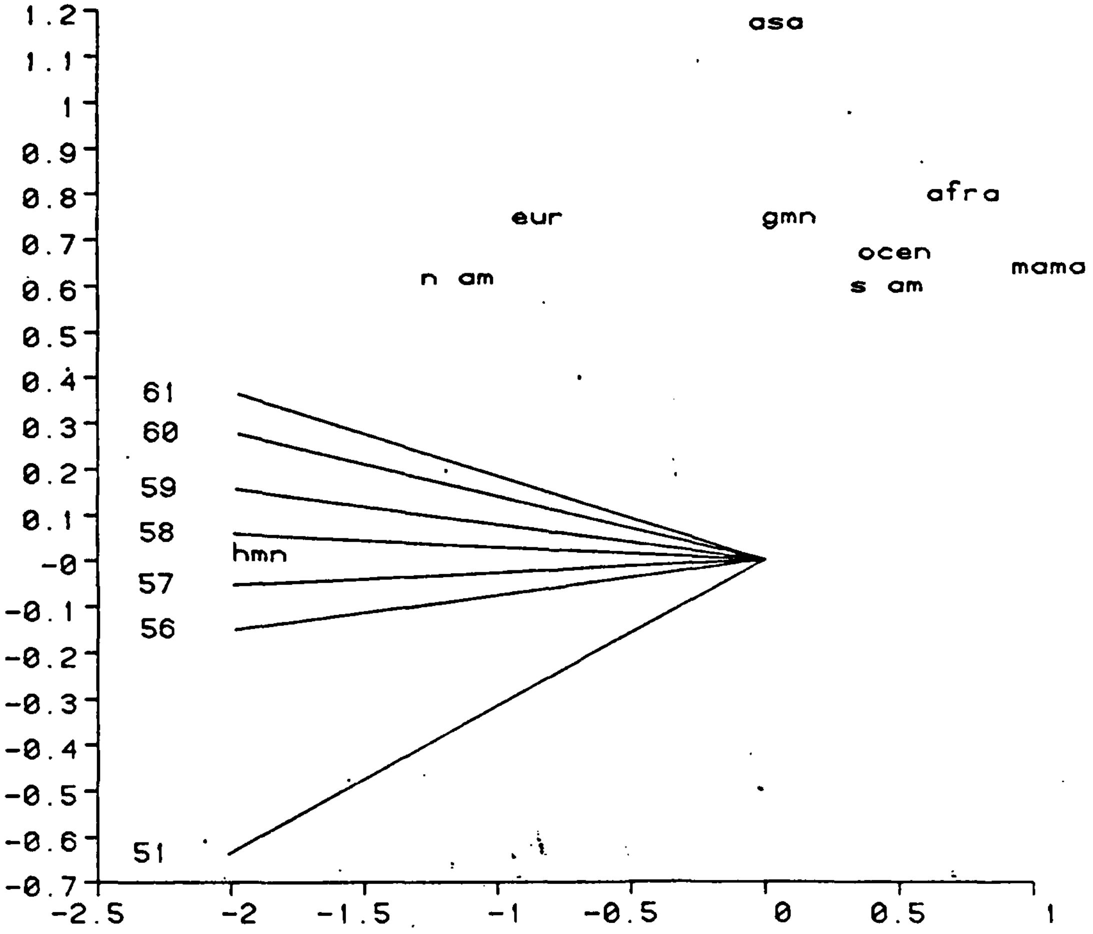


Fig. 9. Biplot of log counts of telephones, by continent and year.

TABLE III A. Least Squares Fit to Logs of Telephone Data.

	Regression Coefficients	
	Intercept	Slope
N.Amer.	10.63	0.063
Eur.	9.86	0.076
Asia	7.80	0.116
S.Amer.	7.45	0.063
Oceania	7.39	0.063
Africa	6.79	0.076
MidAmer.	6.25	0.063
		GOF = 0.9997

TABLE III B. Least Squares Fit to Residuals for Years 1956-61.

	Intercept	Slope
N.Amer.	0.0567	-0.0075
Eur.	0.0207	-0.0033
Asia	-0.1751	0.0197
S.Amer.	0.0887	-0.0103
Oceania	0.0089	-0.0002
Africa	0.0607	-0.0059
MidAmer.	-0.1492	0.0174
		GOF = 0.9999

Table III A shows the results of our first least squares fit. The goodness-of-fit, expressed as the sum of squared residuals divided by the sum of squared deviations from the overall mean, is very good (99.97%), and we are tempted to stop here, or perhaps fit a model with fewer intercept parameters. (This fit is similar to the fit displayed in Exhibit 14 A--Chapter 12--of Tukey's book, if one improves it slightly by adjusting the row slope for Mid-America to be equal to 30.)

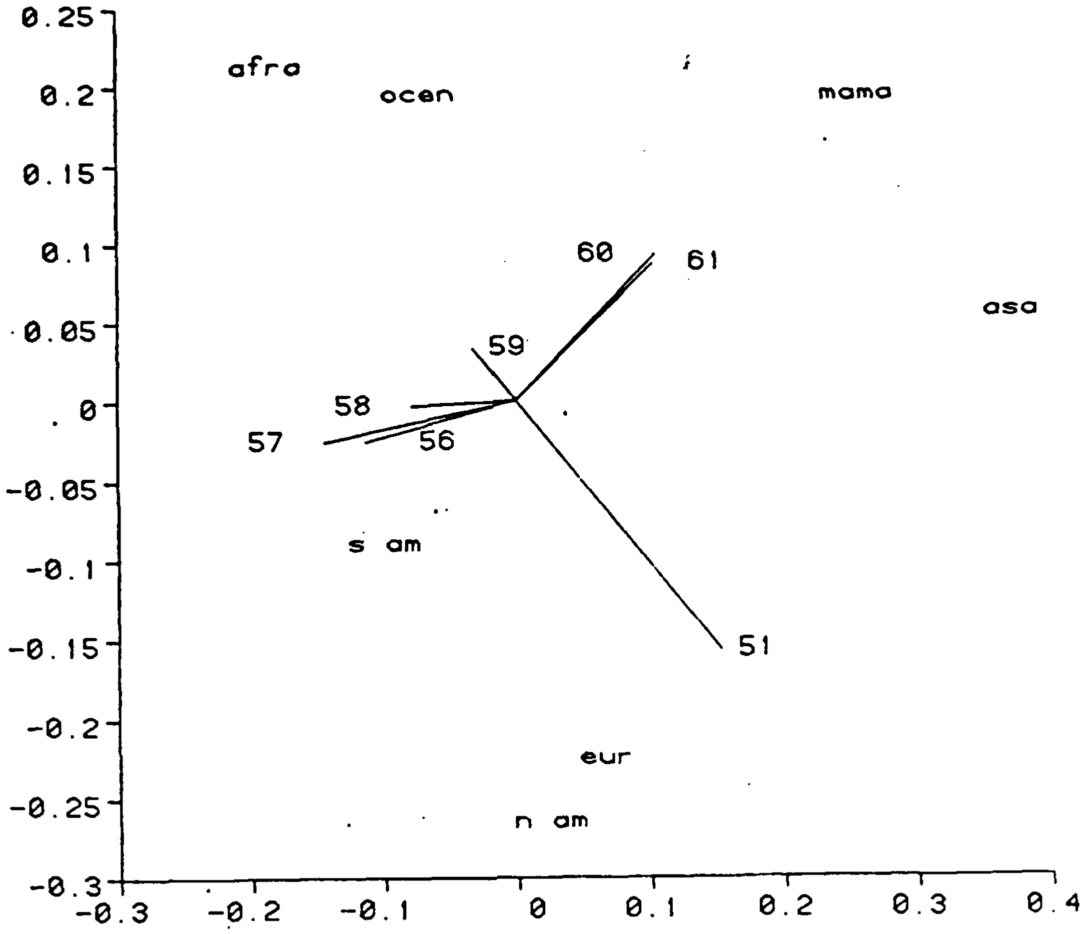


Fig. 10. Biplot of residuals from first fit to log telephone counts.

In Figure 10 we take a second look and biplot the residuals from our first fit. This biplot still shows a trace of collinearity in the column markers from 1956 or 1957 to 1961, and that would diagnose a rows regression model for a subtable. Table III B shows the estimates of the linear regression parameters (one line for each continent) for the 1956-61 subtable. We see that, for Asia and MidAmerica, our first fit overestimated the rates of increase in numbers of telephones from 1951 to 1956, and underestimated them afterwards. The opposite is true for the remaining continents. In general, the year 1951 probably had too great an influence on our first fit. The extra fit makes a small improvement in goodness-of-fit. However, we consider this less important than the extra information we have obtained about time changes in telephone acquisition.

A biplot of the residuals from this additional fit, shown in Figure 11, reveals much less structure than the previous one, although some regularity remains. In this respect one is reminded of the famous "vapor pressure of water" example (Tukey, 1977, Chapter 6) in which definite structure remains in the residuals even after a clearly diagnosed fit.

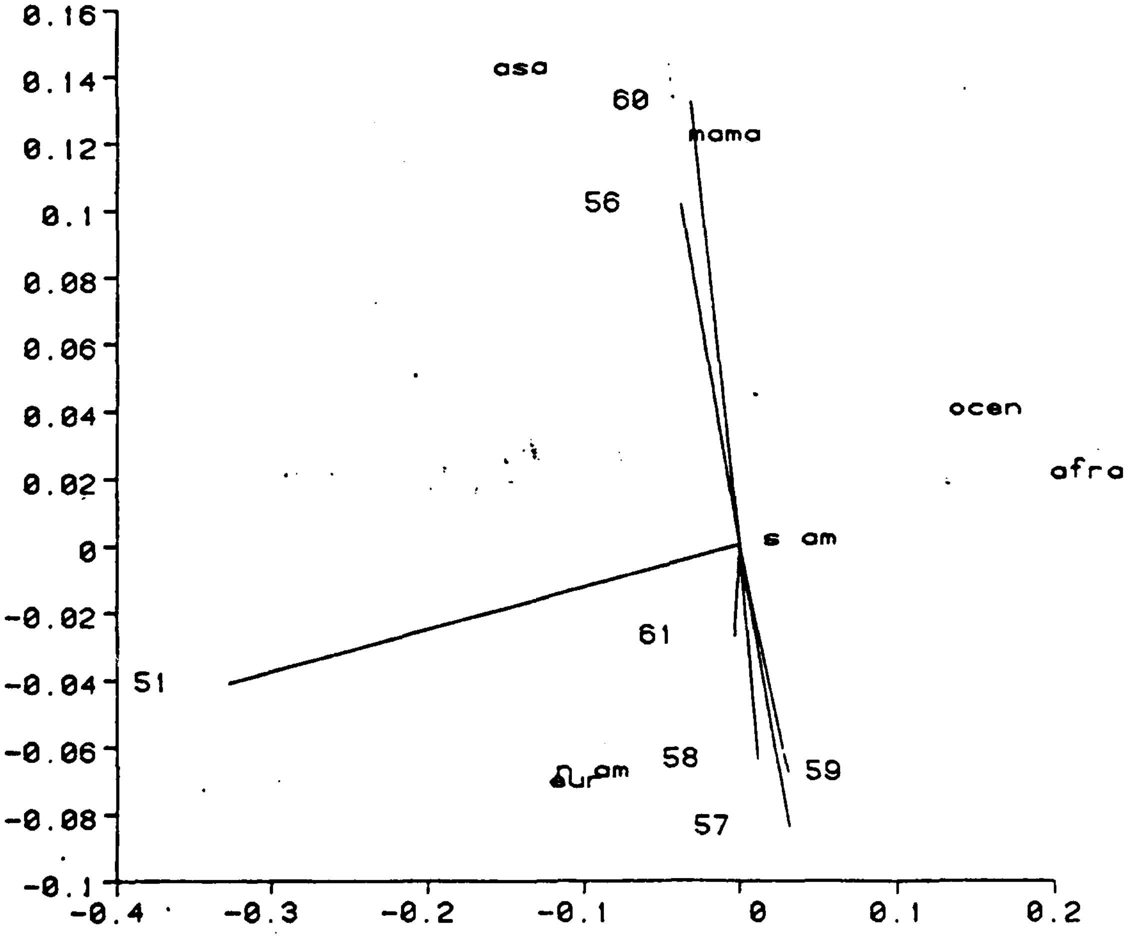


Fig. 11. Biplot of residuals from extra fit to years 1956 to 1961.

Our next example concerns the data in Table IV, which are from an experiment for measuring the sensitivity of several

TABLE IV A. Finger Limens.[a]

			Rates			
	Initial Weights	Persons	r_a	r_b	r_c	r_d
Raw	W_1	K	39	85	101	151
		L	16	32	43	63
		M	18	31	42	58
	W_4	K	31	55	84	124
		L	12	22	33	51
		M	13	26	38	55
	W_7	K	26	56	70	98
		L	12	20	30	37
		M	14	26	40	46

[a] From J. W. Tukey, EDA, Chapter 13; originally P.O. Johnson, Statistical Methods in Research, Prentice-Hall, New York, 1949, Table 89.

TABLE IV B. Logs of Finger Limens.

		Rates			
Initial Weights	Persons	r_a	r_b	r_c	r_d
W_1	K	3.664	4.443	4.615	5.017
	L	2.773	3.466	3.761	4.143
	M	2.890	3.434	3.738	4.060
W_4	K	3.434	4.007	4.431	4.820
	L	2.485	3.091	3.497	3.932
	M	2.565	3.258	3.638	4.007
W_7	K	3.258	4.025	4.248	4.585
	L	2.485	2.996	3.401	3.611
	M	2.639	3.258	3.689	3.829

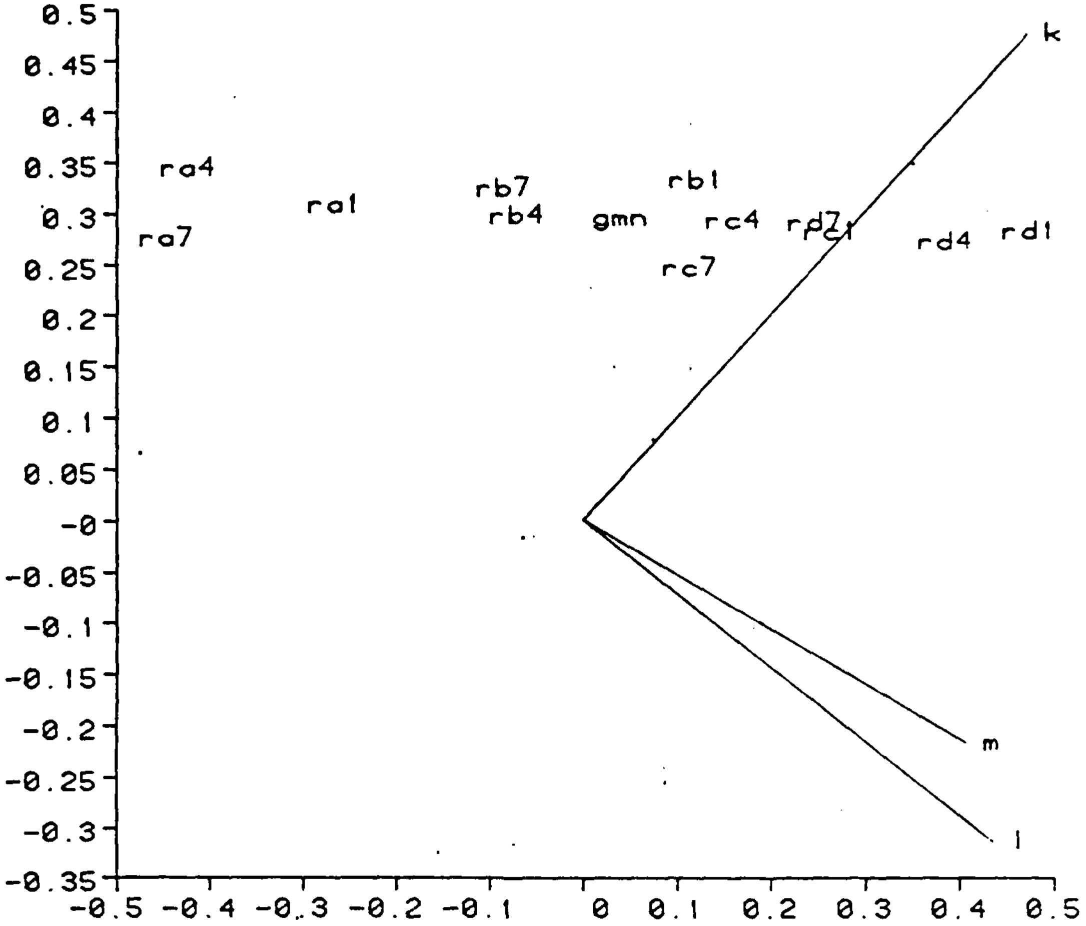


Fig. 12. Biplot of logs of finger limens: rates and weights vs. individuals.

individuals to changes in pull, and are analyzed in Chapter 13 of Tukey's (1977) book. The table shows data for individuals K, L and M, for different initial steady pulls W_1, W_4, W_7, which are referred to as "weights", and different rates of increase of the pull, r_a,r_b, r_c and r_d. We follow Tukey's suggestion and analyze the logs shown in Table IVB.

The first problem with displaying these data is that they are in a three-way layout. As the biplot is a matrix, i.e., two-way, display, it can be applied to these data only if two of the three classifications are crossed either in the rows or in the columns of a matrix--as in Table IV in which weights and individuals are crossed in the rows. There are two other ways of crossing in the rows (there are also three transpositions with crossing in the columns--but their biplots do not differ from the previous three). It will be instructive to look at all three biplots. (Kester (1979) has considered biplot display of such three- and higher-way layouts.)

Figure 12 shows the biplot with individuals in the columns and the rates and weights crossed in the rows. At first it is a little difficult to see any pattern because there are too many row markers. But if one pencils in lines to join the three weights for each rate, a clear pattern emerges. The average of the r_a markers is farthest to the left, then the average of the r_b markers, then that for r_c and finally, the average for r_d--and those four averages are more or less collinear. Furthermore, this line of averages is approximately at right angles to a line through the three markers for individuals. Thus the rate classification appears orthogonal to that by individuals. According to the rules in Figure 7, this suggests additivity between rates and individuals.

From this figure, one can also infer the relative sizes of the differences between rates and weights. In this biplot the positive direction is to the right, as the arrows point there (recall the inner-product construction). The order of the rates in that direction is $r_a < r_b < r_c < r_d$. The order of weights is $W_7 < W_4 < W_1$, but the average differences between the weights are much smaller than those between the rates. We will not discuss such comparisons in detail but rather focus on model diagnosis.

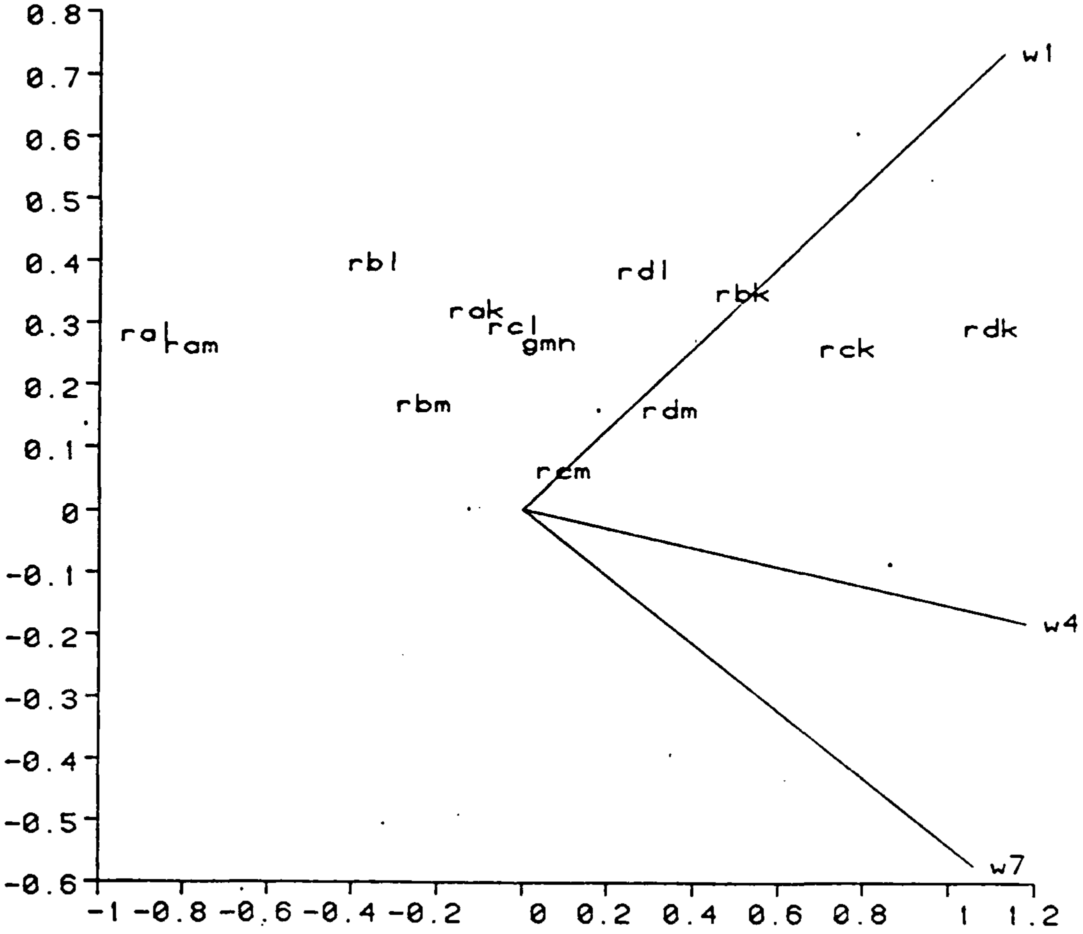


Fig. 13. Biplot of logs of finger limens: rates and individuals vs. weights.

Figure 13 has individuals crossed with rates in the rows. The column markers for weights are pretty much collinear. The row markers seem a bit messy but if one looks at them carefully or draws a few lines, one sees that for each individual the rates are close to a straight line orthogonal to the direction of the weights' line. The diagnosis would therefore be that rates are additive with weights.

From Figure 12, we have rates additive with individuals; from Figure 13, we have rates additive with weights. Together, these diagnoses indicate a model $y_{krw} = \alpha_r + \Theta_{kw}$, in which the variable y is indexed by individual k, rate r, and weight w. The model has a rate effect α_r which is additive to a joint weight-individual effect Θ_{kw}, which allows weight-individual interaction. Indeed, in Figure 13, K, L and M are not collinear, so there is no reason to expect individuals to be additive with weights. It is easy to see that the interaction is due mostly to individual M: the weight markers line is seen to be pretty much orthogonal to K averages, but the average of the M markers is not on that line. The interaction is thus seen to consist mainly of individual M's having a relatively large value for weight 7 and a relatively small value for weight 1. These finding are very similar to those in the analysis in Tukey's (1977) book.

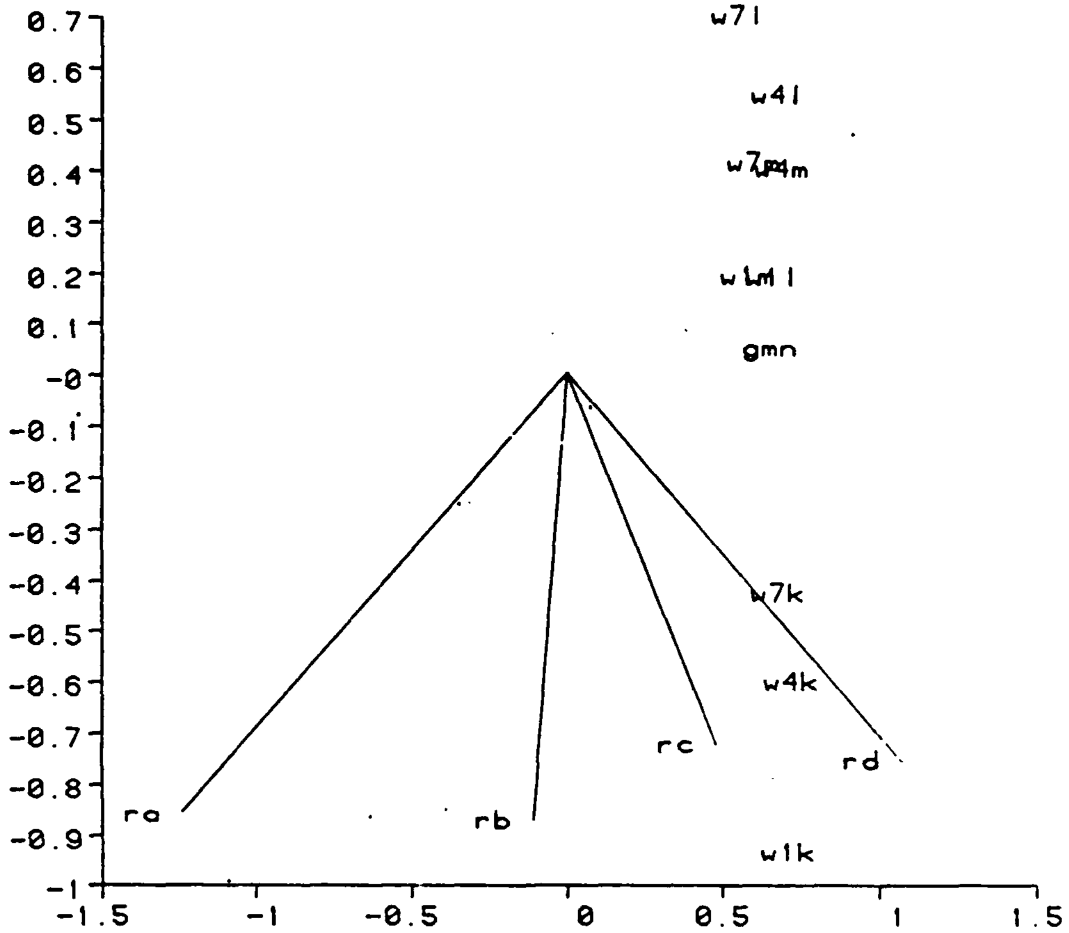


Fig. 14. Biplot of logs of finger limens: weights and individuals vs. rates.

The third biplot is shown in Figure 14. This shows a rather more striking feature than either of the previous two. The rates are represented by fairly collinear column markers; individuals and weights are represented together by collinear row markers at a nearly right angle to the line which roughly fits the rate markers. The most striking feature of this figure is that weights and individuals markers are jointly collinear. By the first rule of Figure 7, the model is diagnosed as $y_{krw} = \alpha_r + \beta_r \Theta_{kw}$--a regression of the rates onto

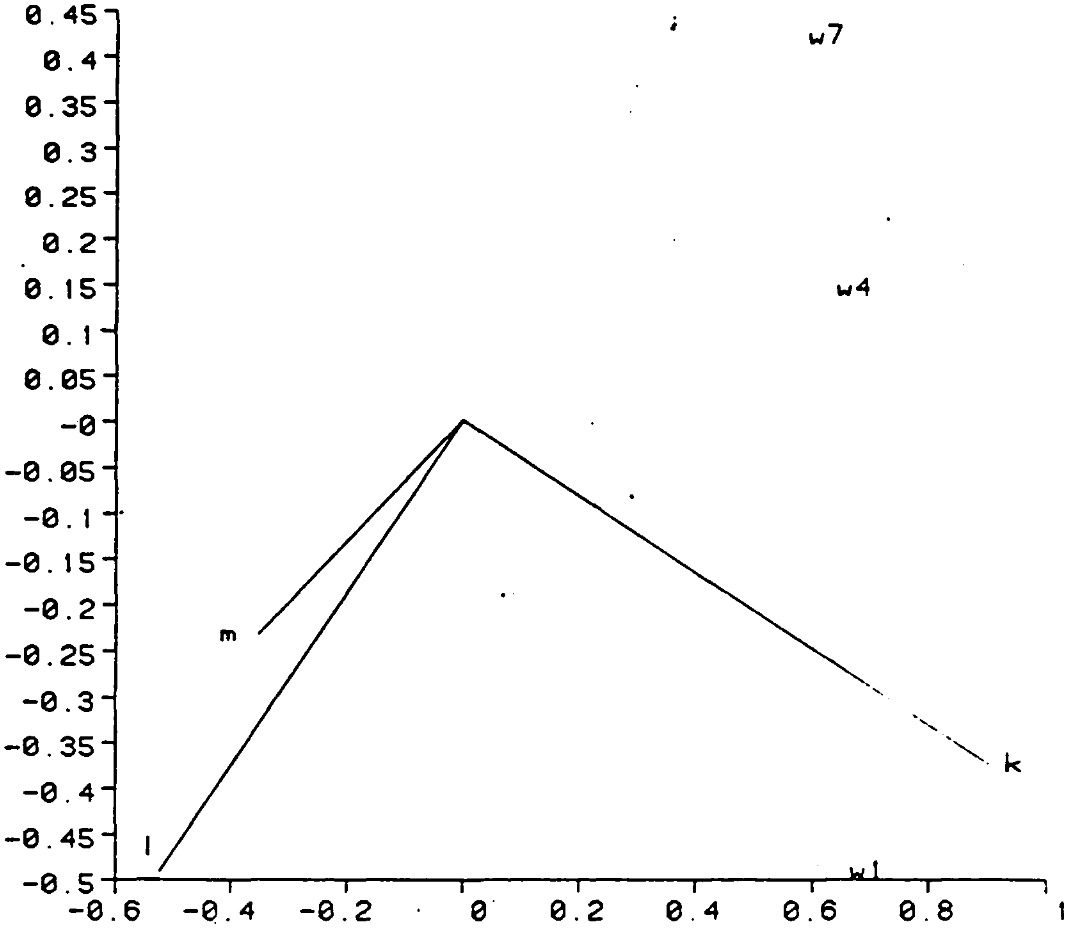


Fig. 15. Biplot of logs of finger limens: weights vs. individuals (averaged over rates).

the weights--individuals combinations: Θ_{kw} is a weight-individual effect and β_r, α_r are the slope and intercept for rate r. Note that this model is a little more general than the one we had before, which we could obtain by setting $\beta_r = 1$ for all r in the present model.

Since the form of the individuals-weights interaction is still uncertain, we consider one more biplot. Figure 15 shows the individuals-weights responses averaged over the four rates.

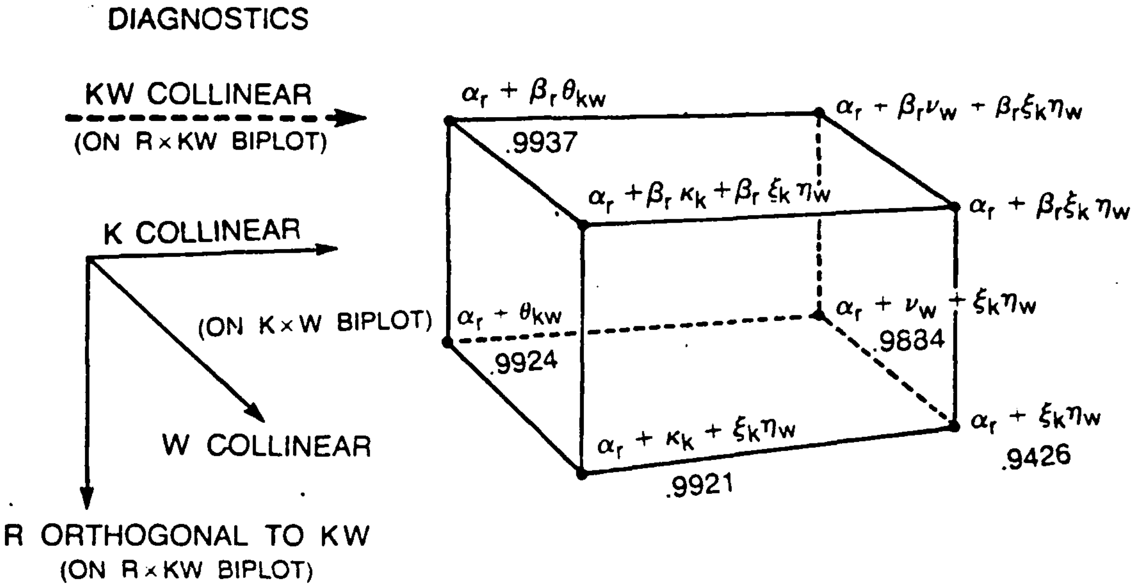

DIAGNOSTICS
KW COLLINEAR
(ON R×KW BIPLOT)
K COLLINEAR
(ON K×W BIPLOT)
W COLLINEAR
R ORTHOGONAL TO KW
(ON R×KW BIPLOT)
$\alpha_r + \beta_r\theta_{kw}$
$\alpha_r + \beta_r\nu_w + \beta_r\xi_k\eta_w$
.9937
$\alpha_r + \beta_r\kappa_k + \beta_r\xi_k\eta_w$
$\alpha_r + \beta_r\xi_k\eta_w$
$\alpha_r + \theta_{kw}$
$\alpha_r + \nu_w + \xi_k\eta_w$
.9884
.9924
$\alpha_r + \xi_k\eta_w$
$\alpha_r + \kappa_k + \xi_k\eta_w$
.9426
.9921

Fig. 16. A summary of models for logs of finger limens.

Here again individual K (a male) is seen to be very different from individuals L and M (females) and the weights show nice collinearity. The second row of the diagnostic table (Figure 7) applies, so the model for θ_{kw} is $\theta_{kw} = \kappa_k + \xi_k\eta_w$, a regression of individuals onto weights.

All these models can be pulled together schematically as shown in Figure 16. Each vertex of the cube is identified with a different model, ranging from the most general (and best fitting) at the top left back corner, to the most specific (and least well fitting) at the bottom right front corner. The biplot diagnostics which indicate specific changes in the model are indicated as directions around the cube. Thus, biplot collinearity of KW-markers suggested the general model $\alpha_r + \beta_r\Theta_{kw}$. Orthogonality of the R-markers to the KW-markers diagnosed absence of R vs. (KW) interaction: the downward arrow therefore indicates modelling in which the R effects are additive to (KW) effects, e.g. $\alpha_r + \beta_r\Theta_{kw}$ becomes $\alpha_r + \Theta_{kw}$.

Collinearity of the K-markers (on the K x W biplot) diagnoses a regression of W onto K: the rightward arrow indicates models in which K effects appear only as "regressors", e.g., θ_{kw} becomes $\nu_w + \xi_k\eta_w$ and $\kappa_k + \xi_k\eta_w$ becomes $\xi_k\eta_w$. Similarly, collinearity of the W-markers (on the K x W biplot) diagnoses regression of K onto W: the forward arrow thus indicates that W effects appear only as "regressors".

The original fit of the most general $\alpha_r + \beta_r\theta_{kw}$ model, diagnosed by KW-collinearity, was 0.9937. Additional diagnoses make for more specific models which are more easily interpretable but fit less well. Thus, the diagnosis by orthogonality simplifies the model to $\alpha_r + \theta_{kw}$ while hardly worsening the fit. On the other hand, the K collinearity diagnosis appreciably reduces the fit in this example. A good model to settle on might be $\alpha_r + \kappa_k + \xi_k\eta_w$ which has a goodness-of-fit of 0.9921--this was diagnosed by the W-collinearity alone.

Again, biplot patterns have diagnosed models very similar to those suggested by John Tukey (1977) using pencil-and-paper EDA methods.

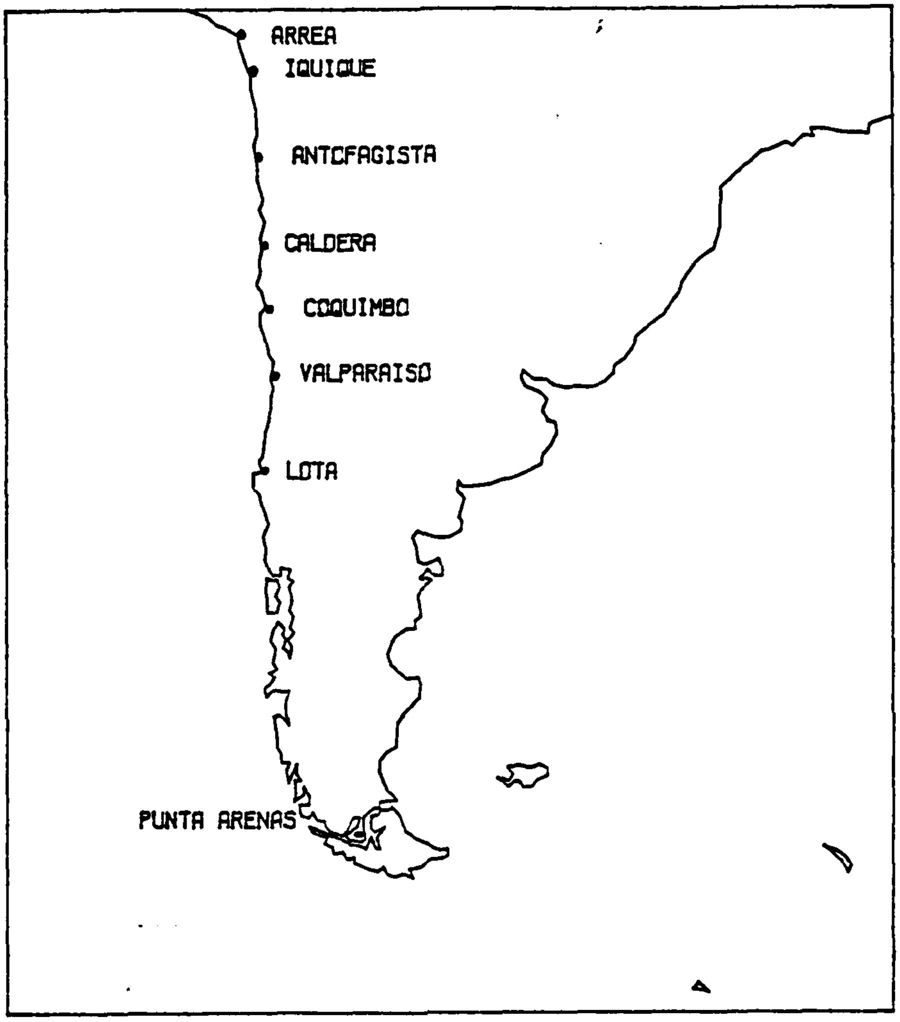


Fig. 17. Seven South American ports.

Our last example deals with ports along the western coast of South America--Figure 17. Ship route distances between these ports are given in Table V A. To begin with, these are arranged in a North-South order so they are a little easier to look at--Table V B. Then the mean distance is subtracted out--Table V C. This makes for a strange kind of

TABLE V A. Shiproute Distances Between S. American Ports.[a]

(distances in sea miles)

	Ant	Arr	Cal	Coq	Iqu	Lota	P A	Val
Ant	0	325	215	396	224	828	1996	576
Arr	325	0	522	702	110	1134	2301	882
Cal	215	522	0	196	420	628	1795	376
Coq	396	702	196	0	602	455	1623	203
Iqu	224	110	420	602	0	1033	2201	782
Lota	828	1134	628	455	1033	0	1191	268
P A	1996	2301	1795	1623	2201	1191	0	1432
Val	576	882	376	203	782	268	1432	0

[a] From J. W. Tukey, EDA, Chapter 11; originally The World Almanac and Book of Facts, New York World-Telgram and Sun.

TABLE V B. N-S Order.

	Arr	Iqu	Ant	Cal	Coq	Val	Lota	P A
Arr	0	110	325	522	702	882	1134	2301
Iqu	110	0	224	420	602	782	1033	2201
Ant	325	224	0	215	396	576	828	1996
Cal	522	420	215	0	196	376	628	1795
Coq	702	602	396	196	0	203	455	1623
Val	882	782	576	376	203	0	268	1432
Lota	1134	1033	828	628	455	268	0	1191
P A	2301	2201	1996	1795	1623	1432	1191	0

TABLE V C. Mean-Centered Data.

	Ant	Arr	Cal	Coq	Iqu	Lota	P A	Val
Ant	-732.	-407.	-517.	-336.	-508.	96.	1264.	-156.
Arr	-407.	-732.	-210.	- 30.	-622.	402.	1569.	150.
Cal	-517.	-210.	-732.	-536.	-312.	-104.	1063.	-356.
Coq	-336.	- 30.	-536.	-732.	-130.	-277.	891.	-529.
Iqu	-508.	-622.	-312.	-130.	-732.	301.	1469.	50.
Lota	96.	402.	-104.	-277.	301.	-732.	459.	-464.
P A	1264.	1569.	1063.	891.	1469.	459.	-732.	700.
Val	-156.	150.	-356.	-529.	50.	-464.	700.	-732.

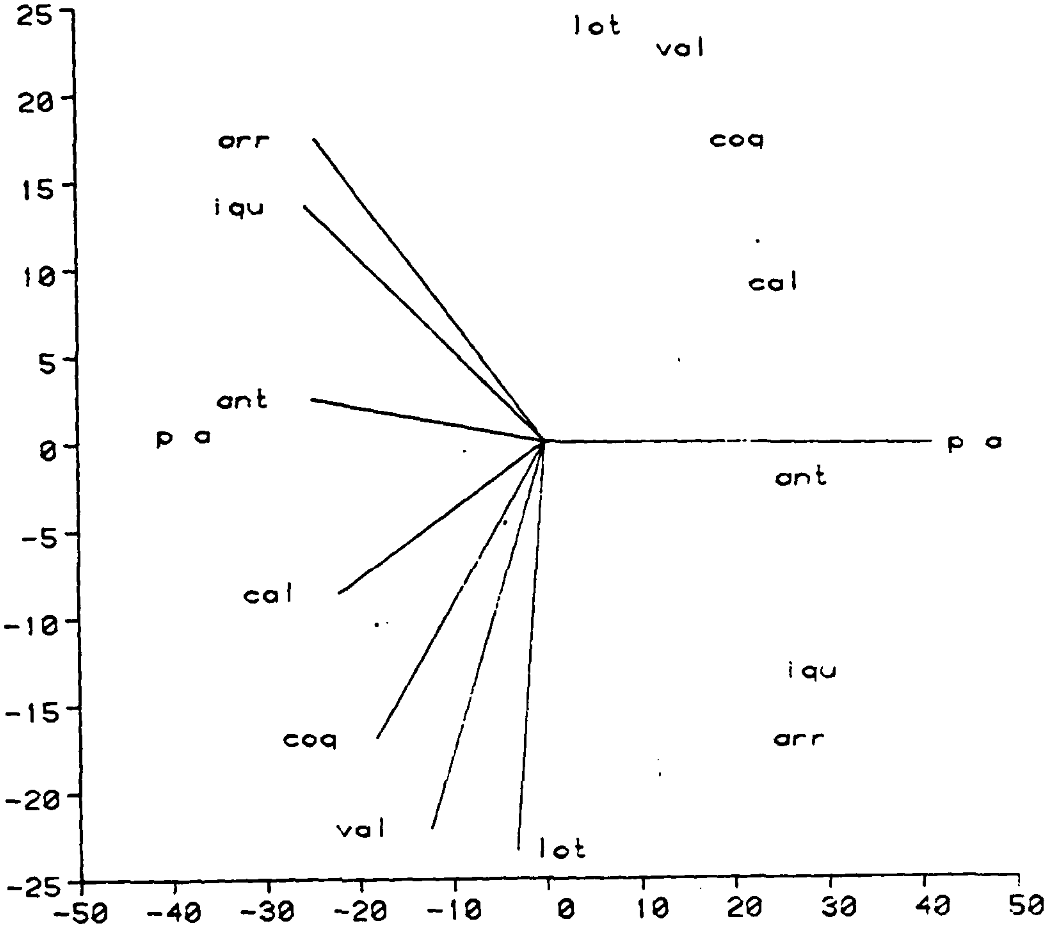


Fig. 18. Biplot of distances between South American ports (mean-centered).

distance, but one that is easier to biplot. Figure 18 shows the biplot of these mean-centered distances. It is immediately evident that the row markers are an exact reflection of the column markers. This is not really surprising since the matrix is symmetric.

It is of interest to consider what is special about biplots of symmetric matrices. Since Y = Y' it follows that in the factorizations AB' = BA'. Thus, one may wonder whether A = B or A = -B, or what else may account for this symmetry. If A = B, one may display an ordinary biplot of factorization AA' in which row markers coincide with column markers: one set of markers suffices. On the other hand, if A = -B, the display of factorization -AA' leads to a biplot like Figure 18 in which the row markers are reflections $-\underline{a}_i$ of the corresponding column markers $+\underline{a}_i$. This redundancy on the biplot can be eliminiated by displaying markers $\underline{a}_i$ along imaginary axes--an imaginary biplot, so to say. One may also achieve this by displaying only the $\underline{a}_i$'s--and not displaying their negatives--but defining the representation by means of the negative inner product, i.e., $y_{i,e} = -\underline{a}_i'\underline{a}_e$. Geometrically, this can be visualized exactly as an ordinary inner-product except that the sign is negative when the $\underline{a}_e$ projection onto $\underline{a}_i$ is in the direction of $\underline{a}_i$ and positive if it is in the opposite direction. That is quite easy to use in practice. Algebraically, we should be thinking of factorization ΔΔ' where the e-th row of Δ is $\underline{\delta}_e' = i\underline{a}_e'$ so that $\underline{\delta}_e'\underline{\delta}_g = (i\underline{a}_e)'(i\underline{a}_g) = -\underline{a}_e'\underline{a}_g$. And the representation of $\underline{\delta}$'s along two imaginary axes looks exactly like that of the $\underline{a}$'s but the imaginary units on the axes produce negative inner-products. (See also Gabriel, 1978, for a biplot with one real and one imaginary axis.)

In the present example of distances the representations and inner-product relations are shown in Figure 19. Large distances, small distances and average distances translate to mean-centered distances above zero, below zero and about zero, respectively.

Distance e to g	Mean-Centered Distance $y_{e,g}$	$\underline{a}_e'\underline{b}_g = \underline{\delta}_e'\underline{\delta}_g$	$\underline{a}_e'\underline{a}_g$	If a's lengths constant	
				angle $(\underline{a}_e,\underline{a}_g)$	$\underline{a}_e$, $\underline{a}_g$
Large	>0	>0	<0	$(\pi/2,\pi]$	distant
Average	0	0	0	$\pi/2$	orthogonal
Small	<0	<0	>0	$[0,\pi/2)$	close

Fig. 19: On the biplot representation of geographic distances.

In the ordinary biplot representation these $(y_{e,g} - \bar{y})$'s are properly represented by $\underline{a}_e'\underline{b}_g$'s. But that is equal to $\underline{\delta}_e'\underline{\delta}_g = (i\underline{a}_e)'(i\underline{a}_g)$. If one leaves out the i and keeps only the real part, then the sign changes, i.e., $\underline{a}_e'\underline{a}_g = -\underline{\delta}_e'\underline{\delta}_g$. Thus for large distances $y_{e,g}$, the inner-product $\underline{a}_e'\underline{a}_g$ would be negative; for small distances, it would be positive; and for average distances, it would be about zero. Moreover, if the $\underline{a}$'s are all of equal length, the inner-product is simply the cosine and varies as the distance between the $\underline{a}$ points. What this means is that when $\underline{a}_e'\underline{a}_g$ is large, there is an obtuse angle. When $\underline{a}_e'\underline{a}_g$ is zero, there will be a 90° angle, and when $\underline{a}_e'\underline{a}_g$ is positive, the angle will be acute. In terms of distances between the $\underline{a}$'s these correspond to large, average and small distances, respectively. This relation between the $\underline{a}$'s thus turns out to be the same as the relation of the original distances $y_{e,g}$. That is why it was convenient to mean-center these distances: the representation along two imaginary axes turned out to be much the same as the original pattern of distances. With this in mind, it is enough to plot the $\underline{a}$'s, which is equivalent to plotting the $\underline{\delta}$'s along imaginary axes, and to consider only the column markers on Figure 18.

This example shows that on occasion one can make use of imaginary biplots for good display of data. John Tukey's (1977) treatment was quite different. Instead of looking at the data, he first tried a model which was intuitively appealing. He postulated that the distance between port e and port g is the sum of (1) a local distance l_e from port e to the shipping lane, (2) a distance $p_{e,g}$ along the shipping lane, and (3) a distance l_g from the lane into port g. He further postulated that shipping lane distances $p_{e,g}$ are simply additive, thus $p_{1,4} = p_{1,2} + p_{2,3} + p_{3,4}$, etc. This model is shown in Figure 20. If one takes

Distance	Port 1	Port 2	Port 3
Port 1	0	$l_1 + p_{1,2} + l_2$	$l_1 + p_{1,2} + p_{2,3} + l_3$
Port 2		0	$l_2 + p_{2,3} + l_3$
Port 3			0

Fig. 20. Tukey's model for nautical distances.

any tetrad on one side of the diagonal of this distance matrix, its four points show additivity, e.g., $(l_1 + p_{1,3} + l_3) - (l_1 + p_{1,5} + l_5) - (l_2 + p_{2,3} + l_3) + (l_2 + p_{2,5} + l_5) = 0$. On the other hand, a tetrad across the diagonal does not have zero differences. In other words, Tukey's model has $y_{i,j} - y_{i,g} - y_{e,j} + y_{e,g} = 0$ whenever $i < e < j < g$. What would happen in factorization $Y = AA'$ (or $Y = \Delta\Delta'$) of such a matrix? The above tetrad condition is readily seen to become $(\underline{a}_i - \underline{a}_e)'(\underline{a}_j - \underline{a}_g) = 0$ for $i < e < j < g$. In other words, $\underline{a}_i - \underline{a}_e$ is orthogonal to $\underline{a}_j - \underline{a}_g$ whenever $i < e < j < g$. A display of such a model for eight ports is readily seen to require eight vectors $\underline{a}_1, \ldots, \underline{a}_8$ such that $\underline{a}_1 - \underline{a}_2$, $\underline{a}_3 - \underline{a}_4$, $\underline{a}_5 - \underline{a}_6$ and $\underline{a}_7 - \underline{a}_8$ are mutually orthogonal. These are only part of the orthogonalities postulated by the model, but they already require a seven dimensional space to represent them. Evidently such a model cannot be diagnosed on a biplot which is two-dimensional, nor on a 3D bimodel. It is essentially a higher dimensional model.

We have discussed this model in some detail because it is an example of what a biplot cannot diagnose. We have found the biplot to be good for diagnosing some models which are (close) to being two or three dimensional, but this is a case of a model which the biplot just cannot represent because the model cannot be collapsed into a plane or three-space.

* * * *

Finally, what have we learned from these examples? How does biplot inspection compare with the EDA methods proposed in Tukey's (1977) book? Parenthetically, we want to remark that the issue is not one of pencil-and-paper methods of median polish versus computer fitting by least squares, because EDA methods have been computerized. The issue we are addressing is which method gives more insight into the form of models that fit the data. Our experience suggests that in using the biplot, a few displays suffice to reveal relevant patterns in a pretty striking manner. EDA, on the other hand, requires several stages of median polish, inspection of residuals, modelling, and re-expression, further median polish, etc., until one may diagnose a model. The biplot is more immediate: It allows one to see things at a glance.

EDA may show more detail if one inspects fits and residuals carefully at each stage, but it requires iterative cycles of modelling, fitting, residuals inspection, re-expression and decisions. If one fit is inadequate, another is tried until a model is judged adequate. This is a search by trial and error rather than by a systematic method. Moreover, the decisions on model choice are based at each stage on I.I.I.--inspired inspection of irregularities. Irregularities are provided by data, inspection takes time, but inspiration is something that may be difficult to come by. In summary, the EDA modelling procedure is in general not systematic. (An exception to this is Tukey's diagnostic plot which uses comparison values systematically for diagnosis of models and re-expressions.)

Biplot diagnostics are more systematic and direct. One does not start by guessing a model, but rather displays data and inspects it--the diagnosis is then often immediate. We know what biplot collinearities mean; we know what right angles mean; we know what coplanarity means and we know something about distances. Identification of any of these patterns makes modelling automatic and hence, to a large extent, objective. And yet, there is also an interactive and somewhat subjective aspect to biplot modelling. One may use one's prior knowledge about the subject matter to choose among various patterns apparent on the biplot. Tsianco (1980) and Gabriel saw the ellipse in the temperature data, though they were not looking for it and at that time had no idea of how to use such a pattern. But as they traced the seasonal variation of the monthly temperatures, they were led to the elliptical pattern. Similarly, when we identify subtables with simple patterns, we interact with the data's display. So biplot modelling is partly systematized and yet allows the investigator to interact with his data and look for interesting patterns.

To sum up, we have sought to demonstrate, by the examples of this paper, that the EDA methods presented in Tukey's (1977) "Golden Book" are not the only ones available. Much can also be learned about suitable models, and most of the messy trial and error of EDA can be avoided, by displaying the data in a biplot.

REFERENCES

Bradu, D. and Gabriel, K. R. (1978). "The Biplot as a Diagnostic Tool for Models of Two-Way Tables," Technometrics, 20, 47-68.

Cox. C., Davis, H. T., Wardell, W. M., Calimlim, J. F., and Lasagna, L. (1980). "Use of the Biplot for Graphical Display and Analysis of Multivariate Pain Data in Clinical Analgesic Trials," Submitted to Controlled Clinical Trials.

Gabriel, K. R. (1971). "The Biplot Graphic Display of Matrices with Application to Principal Component Analysis," Biometrika, 58, 453-467.

Gabriel, K. R. (1978). "The Complex Correlational Biplot," Theory Construction and Data Analysis in the Social Sciences (S. Shye, ed.) San Francisco: Jossey-Bass, 350-370.

Gabriel, K. R. (1980). "Biplot Display of Multivariate Matrices for Inspection of Data and Diagnoses," Interpreting Multivariate Data. (V. Barnett, ed.) London: Wiley (To appear).

Gabriel, K. R., Rave, G. and Weber, E. (1976). "Graphische Darstelling von Matrizen durch das Biplot," EDV in Medizin und Biologie, 7, No. 1, 1-15.

Gabriel, K. R., and Zamir, S. (1979). "Lower Rank Approximation of Matrices by Least Squares with any Choice of Weights," Technometrics, 21, 489-498.

Haber, M. (1975). "The Singular Value Decomposition of Random Matrices," Ph. D. thesis at Hebrew University, Jerusalem.

Householder, A. S., and Young, G. (1938). "Matrix Approximation and Latent Roots," Am. Math. Monthly, 45, 165-171.

Kester, Nancy K. (1979). "Diagnosing and Fitting Concurrent and Related Models for Two-Way and Higher-Way Layouts," Ph. D. thesis at University of Rochester, Rochester, New York.

McNeil, D. R., and Tukey, J. W. (1975). "Higher-Order diagnosis of Two-Way Tables," Biometrics, 31, 487-510.

Mosteller, F., and Tukey, J. W. (1977). Data Analysis and Regression, Reading, Massachusetts., Addison-Wesley.

Tsianco, M. C. (1980). "Use of Biplots and 3D-Bimodels in Diagnosing Models for Two-Way Tables," Ph. D. thesis at University of Rochester, Rochester, New York.

Tukey, J. W. (1977). Exploratory Data Analysis, Reading, Massachusetts., Addison-Wesley.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO. AD-A114 172	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Some Comparisons of Biplot Display and Pencil-and-Paper E.D.A. Methods		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Christopher Cox and K. Ruben Gabriel		8. CONTRACT OR GRANT NUMBER(s) N-0014-80-C-0387
9. PERFORMING ORGANIZATION NAME AND ADDRESS Division of Biostatistics University of Rochester Medical School Rochester, NY 14642		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Arlington, Virginia 22217		12. REPORT DATE June, 1980
		13. NUMBER OF PAGES 37
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Multivariate data may be explored by a variety of methods. This paper considers some examples of alternative analyses by biplot display and by Tukey's pencil-and-paper EDA methods. It suggests that in using the biplot, a few displays usually suffice to reveal patterns in a pretty striking manner. When using EDA, on the other hand, one may require several stages of median polish, inspection of residuals, modelling, and re-expression.

(over)

20 continued:

The biplot is more immediate: It allows one to see things at a glance.

ILMED

8