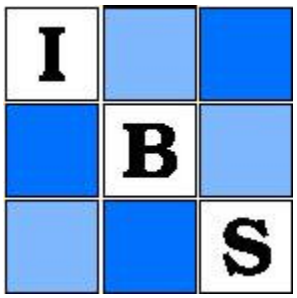

BIOMETRICS *28*, 81–124
March 1972

ROBUST ESTIMATES, RESIDUALS, AND OUTLIER DETECTION WITH MULTIRESPONSE DATA

R. GNANADESIKAN AND J. R. KETTENRING

Bell Telephone Laboratories, Inc., Murray Hill, New Jersey 07974, U. S. A.

SUMMARY

The paper gives an overview of concepts and techniques pertaining to (i) the robust estimation of multivariate location and dispersion; (ii) the analysis of two types of multidimensional residuals—namely those that occur in the context of principal components analysis as well as the more familiar residuals associated with least squares fitting; and (iii) the detection of multiresponse outliers. The emphasis is on methods for informal exploratory analysis and the coverage is both a survey of existing techniques and an attempt to propose, tentatively, some new methodology which needs further investigation and development. Some examples of use of the methods are included.

1. INTRODUCTION

A major objective of statistical data analysis is to extract and explicate the informational content of a body of data. Techniques addressed to this objective involve *summarization*—perhaps in terms of a statistic (e.g. a correlation coefficient) which is undergirded by some tightly specified model, or perhaps in terms of a simple plot (e.g. a scatter plot). It would be very desirable if the methods were also useful for *exposure*—i.e. the presentation of analyses so as to facilitate the detection of not only anticipated but also unanticipated characteristics of the data. The role and value of the twin-pronged process of summarization and exposure in data analysis have been discussed by Tukey and Wilk [1966] and by Gnanadesikan and Wilk [1969].

Statistical methodology for summarization has often been developed without specific attention to the exposure value of the techniques. Even when there has been an awareness of the possible inadequacy or inappropriateness of certain assumptions, the goal of summarization has occasionally been artificially separated from the objective of exposure. For instance, in the routine use of certain robust estimates of location which are designed to be insensitive to outliers, one may overlook the desirability of having techniques for delineating outliers if any exist in a particular body of data. It is clear, however, that from a data-analysis standpoint, the goals of summarization and exposure ought to be viewed as two sides of the same coin.

Statistical estimation of parameters from data is a concern of summarization and, for the uniresponse situation, problems and methods of robust estimation have received considerable attention (cf., for example, Tukey

[1960], Hodges and Lehmann [1963], Huber [1964], and Andrews *et al.* [1971a]). The univariate work has been concerned largely with the problem of obtaining and studying estimates of location which are insensitive to outliers. Also, for comparisons of relative efficiencies and other formal properties of the proposed estimates, the distributions considered as alternatives to the normal have almost always been symmetrical (cf., however, Jaeckel [1971]) with heavier tails than the normal. Thus, even in the univariate case, little attention has been given to the robust estimation of higher order parameters, such as scale and shape parameters, which may be even more susceptible to effects of outliers (cf., however, Huber [1964] and Hampel [1968]).

One of the most insightful processes for exposure in analysis of uniresponse data is the study of residuals in a variety of ways, especially through plotting. For instance, the work of Terry [1955], Anscombe [1960; 1961], Anscombe and Tukey [1963], Daniel [1960], Cox and Snell [1968], and Andrews [1971] has focused attention on the value of analyzing univariate residuals.

The development of robust estimates, as well as methods for analyzing residuals and for detecting outliers, for multivariate situations is very difficult but probably even more needful than in the uniresponse case. One example of the inherent difficulty of the multivariate case is the interrelated character of outlier effects on location, scale, and orientation—one may not always be able to define or delineate outliers for one of these purposes (e.g. location) without considering some of the others (e.g. dispersion). An iterative approach appears to be an integral part of the requirements for the multivariate situation. A general point is that multiresponse models or modes of summarization are inherently more complex, and ways in which the data can depart from the model are infinitely more varied than in the univariate case. Consequently, it is all the more essential to have informal, informative summarization and exposure procedures.

This paper is intended to be an expository overview of concepts and techniques pertaining to multiresponse robust estimation and to detection of multivariate maverick observations. In part it is a survey of existing work. It is, however, also concerned with some new methods. Many of these are proposed tentatively, as possibly useful methods, and they need further theoretical investigation as well as more extensive use. The major concern is with informal exploratory techniques and not formal confirmatory ones such as tests of specific hypotheses.

Specifically, section 2 of the paper discusses robust estimation of multiresponse location and dispersion. Some preliminary findings of a limited Monte Carlo study of the estimates are also included. Section 3 is concerned with methods for analyzing multidimensional residuals. Two types of residuals are defined and methods for studying them are discussed. Section 3.1 discusses residuals associated with principal components analysis while section 3.2 is concerned with the usual type of residuals from least squares fitting. Detection of multivariate outliers is further discussed in section 4, and section 5 consists of concluding comments.

2. MULTIRESPONSE ROBUST ESTIMATORS

In this section the emphasis is on techniques of developing summary statistics which are robust to possible outliers in multiresponse data. The estimation of only relatively simple characteristics, viz. location and dispersion, is considered and we are concerned (exactly as in the univariate case) with protection against outliers, which are considered to be in the 'tails' of the data distribution. The possibilities of other types of outliers, however, do exist in multiresponse data and section 4 discusses the issues of multivariate outliers more fully.

The problem of robust estimation of multivariate location has received some attention in the literature (cf. Mood [1941], Bickel [1964], and Gentleman [1965]) and section 2.1 is, in part, a collation of estimators which have been proposed. Questions of robust estimation of multivariate dispersion, namely both scale (i.e. variance) and orientation (i.e. correlation), have not received much attention. Some issues and approaches relevant to the estimation of dispersion are discussed in section 2.2. Many of the estimators described need further study but do appear to be promising. Preliminary results from a very limited Monte Carlo study are included.

As a convenient notation for drawing the reader's attention, robust estimators are denoted with asterisks while the usual estimators are shown in their conventional forms.

2.1. *Robust estimation of location*

The usual estimator of location is the sample mean vector, $\bar{\mathbf{y}}$. As robust estimators, the following may be considered:

(i) $\mathbf{y}^*_M$, the vector of medians of the observations on each response, as suggested by Mood [1941];

(ii) $\mathbf{y}^*_{HL}$, the vector of Hodges–Lehmann estimators (i.e. the median of averages of pairs of observations) for each response, as proposed and studied by Bickel [1964];

(iii) $\mathbf{y}^*_{T(\alpha)}$, the vector of α-trimmed means (i.e. the mean of the data remaining after omitting a proportion α of the smallest and of the largest observations) for each response.

The use of trimmed and Winsorized means for the univariate case has been advocated by Dixon [1960] and by Tukey [1960; 1962]. The general efficiency characteristics of trimming and Winsorizing appear to be comparable, although Hampel [1968] demonstrates that the 'influence' of extreme observations is greater on the Winsorized than on the trimmed means. At any rate, the present paper does not pursue the use of a vector of Winsorized means in any detail. Also, the work of Gentleman [1965] on estimating multivariate location by minimizing pth power deviations ($1 < p < 2$) is not described and discussed. For present purposes, very simple extensions are the only ones considered.

All of the estimators mentioned above are just vectors of univariate estimators which can be obtained by analyzing the observations on each

response separately. However, the procedure suggested by Gentleman [1965], which is also known to belong to the general class of robust estimators proposed by Huber [1964], is one that is multivariate in character in that the analysis involves a combined manipulation of the observations on the different responses.

One issue, which has been raised by Bickel [1964] in connection with the first two of the above-mentioned estimators, is the lack of affine commutativity of the robust estimators of multivariate location as against the usual mean vector which does possess this property. (A location estimator will be called *affine commutative* if the operations of affine transformation and formation of the estimate can be interchanged without affecting the outcome.) The issue involved here may be viewed in terms of the degree of commitment to the coordinate system for the observations. At one extreme, the interest may be confined entirely to the observed variables and, if so, any issue of commutativity would be remote. At the other extreme, one may feel that the location problem is intrinsically affine commutative and insist that all estimators should have this property. As an intermediate position, one may perhaps seek more limited commutativity (e.g. rotational) than the very general affine commutativity. Although none of the robust estimators of multivariate location discussed in this section have either rotational or affine commutativity, some tentative suggestions for moving in that direction are mentioned in the concluding section.

Comparisons of the proposed location estimators were made using 100 Monte Carlo realizations of $\mathbf{Y}$, a $2 \times n$ matrix, the columns of which are a sample of size n from a bivariate normal distribution with zero means, unit variances, and correlation ρ. Independent samples were formed for $n = 20, 40$, and 60 for $\rho = 0.0$ and then transformed by replacing the second row of each $\mathbf{Y}$ with $\rho y_{1j} + (1 - \rho^2)^{\frac{1}{2}} y_{2j}$, $j = 1, \cdots, n$, in order to construct additional data for $\rho = 0.5$ and 0.9. Thus, even though the results for different sample sizes are independent, the comparisons for fixed n are not.

The data were used to compute the Monte Carlo mean vector (cf. Table 1) as well as the elements, var (1), var (2), and cov (1, 2); the associated correlation, corr (1, 2); and the eigenvalues, $c_{\min}$ and $c_{\max}$, of the Monte Carlo covariance matrix (cf. Tables 2a, b, and c) of $\bar{\mathbf{y}}$, $\mathbf{y}^*_M$, $\mathbf{y}^*_{HL}$, $\mathbf{y}^*_{T(.25)}$, and $\mathbf{y}^*_{T(.1)}$, respectively.

Although the scope of the Monte Carlo experiment was not large enough to establish properties such as the efficiencies (viz. the above variance-type functions of the covariance matrix of $\mathbf{y}^*$) to a satisfactory degree of precision, there are a few general indications which can be inferred from these tables: (i) the location estimators are all reasonably unbiased (as expected!); (ii) there are considerable differences among the efficiencies, except for $\mathbf{y}^*_{HL}$ and $\mathbf{y}^*_{T(.1)}$ which behave much alike; (iii) for $\rho = 0.5$ and 0.9 the correlations between the elements of the robust estimators (especially $\mathbf{y}^*_M$) are less than the correlations for the sample mean vector.

To facilitate further comparisons among the robust estimators, three multivariate measures of the relative efficiencies of the $\mathbf{y}^*$'s with respect

to $\bar{\mathbf{y}}$ were computed (cf. Tables 3a, b, and c), based on the information in Tables 2a, b, and c. The measures of efficiency are the square root of the product, the sum, and the square root of the sum of squares of the eigenvalues of the Monte Carlo covariance matrices of $\bar{\mathbf{y}}$ and $\mathbf{y}^*$. The relative efficiencies are then the corresponding ratios of the values for $\bar{\mathbf{y}}$ to the values for $\mathbf{y}^*$ which, for brevity, are referred to as det, tr, and tr^2, respectively.

Bickel [1964] selected det as a measure of asymptotic relative efficiency in his study of $\mathbf{y}^*_M$ and $\mathbf{y}^*_{HL}$. His results indicate that, for large n and for $\rho = 0.0$, 0.5, and 0.9, det = 0.64, 0.59, and 0.40 for $\mathbf{y}^*_M$ and, for $\mathbf{y}^*_{HL}$, det = 0.96, 0.94, and 0.92.

In contrast to det, which is quite sensitive to the correlations between the elements of $\bar{\mathbf{y}}$ and of $\mathbf{y}^*$, tr is independent of these correlations and hence behaves like the univariate relative efficiency when the variables have equal variances. The third measure also depends on the correlation but, unlike det, tr^2 is not susceptible to difficulty as the correlations approach unity.

The departures of the Monte Carlo results from the above asymptotic results may be attributed to the limitations of the present study and to the effect of the finite sample sizes. Nevertheless, it is quite apparent by any of the three measures that $\mathbf{y}^*_M$ is the least efficient, that the bivariate midmean $\mathbf{y}^*_{T(.25)}$ is somewhat better, and that $\mathbf{y}^*_{HL}$ and $\mathbf{y}^*_{T(.1)}$ are very similar and the most efficient of the lot.

TABLE 1

MONTE CARLO MEANS OF LOCATION ESTIMATES†

		$\bar{x}$	$\underset{\sim}{y}^*_M$	$\underset{\sim}{y}^*_{HL}$	$\underset{\sim}{y}^*_{T(.25)}$	$\underset{\sim}{y}^*_{T(.1)}$
n = 20	ρ = 0.0	$\begin{pmatrix}0.0440\\-0.0064\end{pmatrix}$	$\begin{pmatrix}0.0451\\-0.0016\end{pmatrix}$	$\begin{pmatrix}0.0490\\-0.0003\end{pmatrix}$	$\begin{pmatrix}0.0505\\-0.0065\end{pmatrix}$	$\begin{pmatrix}0.0514\\-0.0022\end{pmatrix}$
	ρ = 0.5	0.0165	0.0014	0.0163	0.0016	0.0120
	ρ = 0.9	0.0368	0.0305	0.0386	0.0292	0.0366
n = 40	ρ = 0.0	$\begin{pmatrix}-0.0235\\-0.0041\end{pmatrix}$	$\begin{pmatrix}-0.0225\\-0.0231\end{pmatrix}$	$\begin{pmatrix}-0.0183\\-0.0041\end{pmatrix}$	$\begin{pmatrix}-0.0194\\-0.0067\end{pmatrix}$	$\begin{pmatrix}-0.0180\\-0.0037\end{pmatrix}$
	ρ = 0.5	-0.0153	-0.0285	-0.0204	-0.0260	-0.0163
	ρ = 0.9	-0.0229	-0.0316	-0.0226	-0.0295	-0.0221
n = 60	ρ = 0.0	$\begin{pmatrix}-0.0056\\-0.0061\end{pmatrix}$	$\begin{pmatrix}-0.0051\\-0.0054\end{pmatrix}$	$\begin{pmatrix}-0.0072\\-0.0081\end{pmatrix}$	$\begin{pmatrix}-0.0030\\-0.0071\end{pmatrix}$	$\begin{pmatrix}-0.0072\\-0.0086\end{pmatrix}$
	ρ = 0.5	-0.0081	-0.0077	-0.0074	-0.0068	-0.0078
	ρ = 0.9	-0.0077	0.0026	-0.0070	-0.0040	-0.0055

†The first elements of the estimates do not vary with the correlation and therefore are included only for the case ρ = 0.

TABLE 2a

MONTE CARLO EFFICIENCIES OF LOCATION ESTIMATES†

(a) $n = 20$

		$\bar{y}$	y^*_M	y^*_{HL}	$y^*_{T(.25)}$	$y^*_{T(.1)}$
	var(1)	0.0478	0.0736	0.0515	0.0615	0.0521
	var(2)	0.0502	0.0626	0.0533	0.0552	0.0521
$\rho = 0.0$	cov(1,2)	0.0057	0.0039	0.0068	0.0082	0.0069
	corr(1,2)	0.1173	0.0578	0.1293	0.1410	0.1315
	c_{min}	0.0431	0.0614	0.0456	0.0496	0.0453
	c_{max}	0.0549	0.0748	0.0592	0.0672	0.0590
	var(2)	0.0546	0.0795	0.0584	0.0630	0.0550
	cov(1,2)	0.0289	0.0325	0.0295	0.0317	0.0289
$\rho = 0.5$	corr(1,2)	0.5655	0.4246	0.5369	0.5099	0.5354
	c_{min}	0.0221	0.0439	0.0253	0.0305	0.0250
	c_{max}	0.0803	0.1092	0.0846	0.0940	0.0829
	var(2)	0.0528	0.0717	0.0559	0.0646	0.0551
	cov(1,2)	0.0455	0.0578	0.0478	0.0553	0.0479
$\rho = 0.9$	corr(1,2)	0.9066	0.7949	0.8909	0.8777	0.8948
	c_{min}	0.0047	0.0149	0.0058	0.0077	0.0056
	c_{max}	0.0959	0.1304	0.1015	0.1184	0.1016

†Note: var(1) does not change with the correlation.

(b) $n = 40$

		$\bar{y}$	y^*_M	y^*_{HL}	$y^*_{T(.25)}$	$y^*_{T(.1)}$
	var(1)	0.0213	0.0314	0.0234	0.0261	0.0224
	var(2)	0.0327	0.0384	0.0343	0.0347	0.0332
$\rho = 0.0$	cov(1,2)	0.0031	0.0009	0.0029	0.0022	0.0020
	corr(1,2)	0.1167	0.0266	0.1019	0.0743	0.0732
	c_{min}	0.0205	0.0313	0.0227	0.0019	0.0221
	c_{max}	0.0335	0.0385	0.0351	0.0608	0.0336
	var(2)	0.0325	0.0443	0.0344	0.0357	0.0347
	cov(1,2)	0.0133	0.0144	0.0139	0.0144	0.0136
$\rho = 0.5$	corr(1,2)	0.5057	0.3862	0.4907	0.4709	0.4892
	c_{min}	0.0124	0.0221	0.0139	0.0157	0.0136
	c_{max}	0.0413	0.0536	0.0439	0.0460	0.0435
	var(2)	0.0258	0.0443	0.0289	0.0326	0.0286
	cov(1,2)	0.0205	0.0267	0.0223	0.0243	0.0218
$\rho = 0.9$	corr(1,2)	0.8734	0.7151	0.8596	0.8332	0.8610
	c_{min}	0.0029	0.0104	0.0036	0.0029	0.0035
	c_{max}	0.0441	0.0653	0.0487	0.0587	0.0475

TABLE 2 (Contd.)

(c) $n = 60$

		$\bar{\underset{\sim}{y}}$	$\overset{*}{\underset{\sim}{y}}_M$	$\overset{*}{\underset{\sim}{y}}_{HL}$	$\overset{*}{\underset{\sim}{y}}_{T(.25)}$	$\overset{*}{\underset{\sim}{y}}_{T(.1)}$
	var(1)	0.0147	0.0217	0.0156	0.0175	0.0154
	var(2)	0.0175	0.0266	0.0190	0.0214	0.0190
$\rho = 0.0$	cov(1,2)	0.0003	0.0010	0.0002	0.0004	0.0001
	corr(1,2)	0.0216	0.0409	0.0114	0.0220	0.0087
	c_{min}	0.0147	0.0215	0.0156	0.0175	0.0154
	c_{max}	0.0176	0.0268	0.0190	0.0215	0.0190
	var(2)	0.0171	0.0272	0.0194	0.0224	0.0196
	cov(1,2)	0.0077	0.0078	0.0075	0.0077	0.0074
$\rho = 0.5$	corr(1,2)	0.4821	0.3195	0.4305	0.3883	0.4243
	c_{min}	0.0082	0.0162	0.0098	0.0119	0.0099
	c_{max}	0.0237	0.0327	0.0252	0.0280	0.0252
	var(2)	0.0155	0.0248	0.0165	0.0182	0.0165
	cov(1,2)	0.0134	0.0151	0.0137	0.0145	0.0137
$\rho = 0.9$	corr(1,2)	0.8862	0.6517	0.8558	0.8142	0.8603
	c_{min}	0.0017	0.0081	0.0023	0.0033	0.0022
	c_{max}	0.0285	0.0385	0.0297	0.0324	0.0297

2.2. *Robust estimation of dispersion*

There are at least two aspects of multivariate dispersion, viz. that depending on the scales (i.e. variances) of the responses and that concerned with orientation (i.e. inter-correlations among the responses). For some purposes it may be desirable to consider the robust estimation of each of these aspects separately, while for other purposes a combined view may be in order.

An approach that separates the two aspects has the advantage of using all the available and relevant information for each estimation task, whereas an approach that combines the two would involve retaining only observations (perhaps fewer in number) which pertain to both aspects simultaneously. On the other hand, in many cases, the combined approach may be computationally simpler and more economical. Some suggestions are made for both types of approaches in the subsections which follow.

2.2.1. *Estimation of variance for each response*

The problem of robust estimation of the variance of a univariate population has been considered (cf. Tukey [1960], Johnson and Leone ([1964] section 6.9), and Hampel [1968]) although not as intensively or extensively as the location case. When one leaves the location case, certain conflicting aims seem to emerge in estimating higher order characteristics of a distri-

TABLE 3

MONTE CARLO RELATIVE EFFICIENCIES

(a) $n = 20$

		$\underset{\sim}{y}^*_M$	$\underset{\sim}{y}^*_{HL}$	$\underset{\sim}{y}^*_{T(.25)}$	$\underset{\sim}{y}^*_{T(.1)}$
$\rho = 0.0$	det	0.718	0.936	0.844	0.944
	tr	0.720	0.935	0.840	0.941
	tr^2	0.721	0.934	0.836	0.939
$\rho = 0.5$	det	0.608	0.910	0.787	0.923
	tr	0.669	0.931	0.822	0.948
	tr^2	0.708	0.943	0.843	0.961
$\rho = 0.9$	det	0.481	0.871	0.702	0.887
	tr	0.692	0.937	0.798	0.938
	tr^2	0.731	0.944	0.809	0.944

(b) $n = 40$

		$\underset{\sim}{y}^*_M$	$\underset{\sim}{y}^*_{HL}$	$\underset{\sim}{y}^*_{T(.25)}$	$\underset{\sim}{y}^*_{T(.1)}$
$\rho = 0.0$	det	0.754	0.929	0.873	0.961
	tr	0.773	0.935	0.887	0.969
	tr^2	0.791	0.940	0.901	0.976
$\rho = 0.5$	det	0.659	0.918	0.842	0.932
	tr	0.710	0.930	0.870	0.941
	tr^2	0.744	0.937	0.887	0.946
$\rho = 0.9$	det	0.438	0.859	0.708	0.886
	tr	0.622	0.900	0.802	0.923
	tr^2	0.669	0.906	0.819	0.928

(c) $n = 60$

		$\underset{\sim}{y}^*_M$	$\underset{\sim}{y}^*_{HL}$	$\underset{\sim}{y}^*_{T(.25)}$	$\underset{\sim}{y}^*_{T(.1)}$
$\rho = 0.0$	det	0.669	0.933	0.830	0.938
	tr	0.668	0.932	0.828	0.937
	tr^2	0.666	0.932	0.827	0.935
$\rho = 0.5$	det	0.604	0.887	0.763	0.883
	tr	0.651	0.911	0.798	0.909
	tr^2	0.686	0.927	0.823	0.926
$\rho = 0.9$	det	0.398	0.845	0.676	0.860
	tr	0.649	0.943	0.847	0.946
	tr^2	0.727	0.957	0.877	0.958

bution. Thus for the variance (and maybe even more so for the shape) there is a possible conflict between the desire to protect the estimate from outliers and the fact that the information for estimating the variance relies more heavily on the tails.

This conflict raises certain questions about the routine use of robust estimation procedures for these higher order characteristics, especially in relatively small samples. Thus with a sample of size 10, for instance, the use of a 10% (the minimum possible in this case) trimmed sample to provide a robust estimate may lead to an estimator whose efficiency is unacceptably low when one is close enough to the normal. The main point is that with relatively small and yet reasonable sample sizes, it may be both expedient and wise to study the observations more closely, omitting only clearly indicated outliers or possibly transforming the observations to make them more nearly normally distributed.

The usual unbiased estimator of the variance for the ith response ($i = 1, \cdots, p$) based on n observations may be denoted s_{ii} and a corresponding robust estimator, s_{ii}^*, may be developed by using any of the following three methods:

(i) trimmed variance from an α-trimmed sample, as suggested by Tukey [1960] and further studied by Hampel [1968];

(ii) Winsorized variance from an α-Winsorized sample as suggested by Tukey and McLaughlin [1963];

(iii) the slope of the lower end of a χ_1^2 probability plot of the $\frac{1}{2}n(n-1)$ squared differences between pairs of observations.

The first two methods need an estimate of location and a direct suggestion would be to use a trimmed mean for the trimmed variance and a Winsorized mean for the Winsorized variance. Huber [1970], however, suggests using a trimmed mean for getting the Winsorized variance, and for t-statistic types of considerations associated with the trimmed mean this may be appropriate. But even for using a trimmed mean for the trimmed variance, or a Winsorized mean for the Winsorized variance, because of the considerations mentioned above, it would seem to be advisable to use a smaller proportion of trimming (or Winsorizing) for the variance estimation than for the location estimation in samples even as large as 20.

To obtain unbiased estimates from a trimmed or Winsorized variance, multiplicative constants are needed. These constants are based on moments of order statistics of the normal distribution and an underlying assumption in using these constants is that the 'middle' of the sample is sufficiently normal. Johnson and Leone ([1964] p. 173) give a table of the required constants for small ($n \leq 15$) samples, and tables provided by McLaughlin and Tukey [1961] together with the tabulation by Teichroew [1956] of the expected values of cross-products and squares of normal order statistics may be used for calculating the required constant for samples of sizes up to 20. Unfortunately, asymptotic results do not appear to be adequate at

$n = 20$, and further work is needed on developing the required multiplicative constants for larger values of n.

One advantage of the third method mentioned above is that it does not involve an estimate of location. A second is that the type of adjustment provided by the multiplicative constant in the trimmed and Winsorized variances is contained in the probability plot itself—viz. the abscissa (or quantile axis) is used to scale the ordinate for determining the slope (which would be an estimate of twice the variance). A third advantage is that, by looking at $\frac{1}{2}n(n-1)$ pieces of information (some of which may be redundant because of statistical correlations), the error configuration on the χ_1^2 probability plot may often be indicated more stably than by a normal probability plot of the n observations. A fourth, and perhaps most significant, advantage of the approach is its exposure value in aiding the detection of unanticipated peculiarities in the data. A disadvantage of the technique is that it may not be useful, and may even be misleading, for estimating the variance in circumstances where a large proportion of the observations may be outliers.

The first two methods were tried out using the previously generated data. The results are shown in Table 4 for the trimmed variance in a typical case, $n = 20$ and $\rho = 0.5$. The empirical evidence supports the contention that an acceptable trimming percentage for location estimation (viz. $\alpha = 0.10$) may be too high for variance estimation. The corresponding results for the Winsorized variance are not included here but are similar.

2.2.2. *Estimation of covariance and correlation*

A simple idea for estimating the covariance between two variables Y_1 and Y_2 is based on the identity

$$\text{cov}\,(Y_1, Y_2) = \tfrac{1}{4}[\text{var}\,(Y_1 + Y_2) - \text{var}\,(Y_1 - Y_2)]. \tag{1}$$

One robust estimator, s_{12}^*, of the covariance between Y_1 and Y_2 may, therefore, be obtained from

$$s_{12}^* = \tfrac{1}{4}(\hat{\sigma}_1^{*2} - \hat{\sigma}_2^{*2}), \tag{2}$$

where $\hat{\sigma}_1^{*2}$ and $\hat{\sigma}_2^{*2}$ are robust estimators of the variances of $Y_1 + Y_2$ and

TABLE 4

MONTE CARLO PROPERTIES OF TRIMMED VARIANCE ESTIMATES ($n = 20$, $\rho = 0.5$)

	% Trimmed	M.C.Mean	M.C.Variance	M.C.Rel.Eff.
s_{11}^*	5	1.0380	0.1382	0.865
s_{11}^*	10	1.0264	0.1683	0.711
s_{22}^*	5	1.0326	0.1698	0.813
s_{22}^*	10	1.0079	0.2183	0.632

$Y_1 - Y_2$, respectively, and may be obtained by any of the methods mentioned in section 2.2.1.

Given such a robust estimator of the covariance, a natural way of defining a corresponding robust estimator of the correlation coefficient between Y_1 and Y_2 is

$$r_{12}^* = s_{12}^*/(s_{11}^* s_{22}^*)^{\frac{1}{2}}, \tag{3}$$

where s_{ii}^* is a robust estimator of the variance of the ith response.

Since the robust estimators involved in (2) and (3) are determined with no considerations of satisfying the well-known Cauchy–Schwarz inequality relationship between the covariance and the variances, therefore r_{12}^* as obtained from (3) may not necessarily lie in the admissible range, $[-1, +1]$, for a correlation coefficient. To ensure an estimate of the correlation coefficient in the valid range, while still remaining with the above approach of obtaining the covariance estimate as the difference between two variance estimates, a modification may be suggested. Let $Z_i = Y_i/\sqrt{s_{ii}^*}$ denote the 'standardized' form of Y_i, where s_{ii}^* is a robust estimate of the variance of Y_i. Then define

$$\hat{\rho}_{12}^* = (\hat{\sigma}_3^{*2} - \hat{\sigma}_4^{*2})/(\hat{\sigma}_3^{*2} + \hat{\sigma}_4^{*2}), \tag{4}$$

where now $\hat{\sigma}_3^{*2}$ and $\hat{\sigma}_4^{*2}$ are robust estimators of the variances of $Z_1 + Z_2$ and $Z_1 - Z_2$, respectively. Corresponding to $\hat{\rho}_{12}^*$, which necessarily lies in the range $[-1, +1]$, a covariance estimator may be defined by

$$\hat{\sigma}_{12}^* = \hat{\rho}_{12}^*(s_{11}^* s_{12}^*)^{\frac{1}{2}}. \tag{5}$$

An interesting property of estimating the correlation coefficient by (3) or (4) is that the multiplicative constant, which is required for removing the biases involved in trimmed or Winsorized variances, cancels out by appearing both in the numerator and denominator of the defining equations (3) and (4). Hence, for any sample size, the trimmed (or Winsorized) variances which provide bases for obtaining r_{12}^* and $\hat{\rho}_{12}^*$ can be used directly without any multiplicative constant. This does not, however, imply that r_{12}^* and $\hat{\rho}_{12}^*$ are unbiased estimators of the population correlation.

Again utilizing the data described in section 2.2, Monte Carlo means, variances, and relative efficiencies (with respect to the usual estimators, s_{12} and r_{12}) were computed for the trimmed and Winsorized versions ($\alpha = 0.05$ and 0.10) of s_{12}^* and $\hat{\sigma}_{12}^*$ ($n = 20$ only) and r_{12}^* and $\hat{\rho}_{12}^*$. The results for the trimming approach are displayed in Tables 5a, b, and c.

The figures for the two covariance estimators are much alike, and it appears that the disparity between the relative efficiencies for $\alpha = 0.05$ and $\alpha = 0.10$ increases with ρ. For the correlation estimators, the indications are that for large ρ, $\hat{\rho}_{12}^*$ is more efficient than r_{12}^* but neither is efficient enough when $\alpha = 0.10$ (even $\alpha = 0.05$ may be too high). A few isolated cases of $|r_{12}^*| > 1$ occurred for $\rho = 0.9$, but the frequency was never greater than 3%. The overall results for the Winsorized estimators were not sufficiently different to warrant discussion except for the comment that for $n = 20$,

TABLE 5

MONTE CARLO PROPERTIES OF TRIMMED COVARIANCE AND CORRELATION ESTIMATES

(a) $n = 20$

		M.C.Mean		M.C. Variance		M.C. Rel.Eff.	
		5%	10%	5%	10%	5%	10%
$\rho = 0.0$	s^*_{12}	-0.0018	-0.0221	0.0901	0.1069	0.796	0.671
	$\hat{\sigma}^*_{12}$	0.0013	-0.0222	0.0925	0.1076	0.775	0.666
	r^*_{12}	-0.0028	-0.0263	0.0778	0.1021	0.762	0.580
	$\hat{\rho}^*_{12}$	0.0025	-0.0216	0.0765	0.0906	0.775	0.654
$\rho = 0.5$	s^*_{12}	0.5187	0.5011	0.0914	0.1079	0.890	0.754
	$\hat{\sigma}^*_{12}$	0.5108	0.4787	0.0920	0.1114	0.885	0.731
	r^*_{12}	0.4962	0.4872	0.0414	0.0574	0.812	0.585
	$\hat{\rho}^*_{12}$	0.4870	0.4578	0.0413	0.0527	0.813	0.638
$\rho = 0.9$	s^*_{12}	0.9412	0.9283	0.1175	0.1483	0.923	0.731
	$\hat{\sigma}^*_{12}$	0.9324	0.9107	0.1172	0.1447	0.925	0.749
	r^*_{12}	0.9041	0.9012	0.0048	0.0082	0.553	0.320
	$\hat{\rho}^*_{12}$	0.8945	0.8839	0.0036	0.0057	0.732	0.461

(b) $n = 40$

$\rho = 0.0$	r^*_{12}	-0.0358	-0.0378	0.0297	0.0332	0.946	0.844
	$\hat{\rho}^*_{12}$	-0.0310	-0.0338	0.0285	0.0318	0.985	0.882
$\rho = 0.5$	r^*_{12}	0.4682	0.4673	0.0183	0.0227	0.914	0.734
	$\hat{\rho}^*_{12}$	0.4610	0.4580	0.0177	0.0208	0.946	0.804
$\rho = 0.9$	r^*_{12}	0.8905	0.8871	0.0017	0.0031	0.642	0.355
	$\hat{\rho}^*_{12}$	0.8889	0.8855	0.0013	0.0018	0.829	0.613

TABLE 5 *Continued*

(c) $n = 60$

$\rho = 0.0$	r^*_{12}	0.0025	-0.0005	0.0217	0.0244	0.905	0.803
	$\hat{\rho}^*_{12}$	0.0009	-0.0022	0.0228	0.0245	0.862	0.802
$\rho = 0.5$	r^*_{12}	0.4913	0.4947	0.0152	0.0195	0.786	0.612
	$\hat{\rho}^*_{12}$	0.4925	0.4911	0.0139	0.0167	0.858	0.713
$\rho = 0.9$	r^*_{12}	0.8995	0.8982	0.0012	0.0020	0.664	0.402
	$\hat{\rho}^*_{12}$	0.8966	0.8945	0.0010	0.0014	0.759	0.583

$\rho = 0.9$, and $\alpha = 0.10$ thirteen incidents of $|r^*_{12}| > 1$ were found in the 100 computed correlations.

2.2.3. *Estimation of the covariance matrix*

The usual estimate of the covariance matrix is the sample covariance matrix, **S**. Given robust estimates of the variances and covariances obtained by the methods of the two previous subsections, a direct method of obtaining a robust estimate of the covariance matrix is just to 'put these together' in a matrix. Thus, corresponding to each of the two methods described above (cf. equations (3) and (4)) for obtaining an estimate of the correlation coefficient, a robust estimate of the covariance matrix would be

$$\mathbf{S}^*_a = \mathbf{D}\mathbf{R}^*_a\mathbf{D} \qquad a = 1, 2, \tag{6}$$

where **D** is a diagonal matrix with diagonal elements $\sqrt{s^*_{ii}}$ $(i = 1, \cdots, p)$, $\mathbf{R}^*_1 = (r^*_{ij})$ and $\mathbf{R}^*_2 = (\hat{\rho}^*_{ij})$.

For some purposes of analyzing the multiresponse data, when the underlying distribution is not singular, it may be desirable to have a positive definite estimate of the covariance matrix. For instance, in analyzing the configuration of the sample in terms of the generalized squared distance of the observations from the sample centroid (cf. Example 2 in section 3.2), the inverse of the estimate of the covariance matrix is used.

If the dimensionality, p, does not exceed the number of independent observations (viz. $n - 1$ in the case of an unstructured sample), then the usual estimator, **S**, is positive definite with probability 1. However, neither of the estimators, $\mathbf{S}^*_1$ and $\mathbf{S}^*_2$, defined above is necessarily positive definite. The positive definiteness of these estimators is equivalent to the positive definiteness of the corresponding estimators, $\mathbf{R}^*_1$ and $\mathbf{R}^*_2$, of the correlation matrix; and even though each off-diagonal element of $\mathbf{R}^*_2$ necessarily lies in the range

$[-1, +1]$, this does not necessarily imply positive definiteness of $\mathbf{R}_2^*$, except for the bivariate case.

Some methods for obtaining positive definite robust estimators of covariance matrices are described next. The essential idea underlying all of them is to base the estimate on a 'sufficiently large' number, ν, of the observations (i.e. $\nu > p$) which are, nevertheless, subselected from the total sample so as to make the estimate robust to outliers. A second feature of these estimators is that they are all based on a combined consideration of both scale and orientational aspects, unlike $\mathbf{S}_1^*$ and $\mathbf{S}_2^*$ which were built up from separate considerations of these aspects.

The first method for ensuring a positive definite robust estimator of the covariance matrix is taken from Wilk *et al.* [1962] who were concerned with developing appropriate compounding matrices for a squared distance function employed in an internal comparisons technique suggested by Wilk and Gnanadesikan [1964] for analyzing a collection of single-degree-of-freedom contrast vectors (cf. section 3.2).

The first step in the procedure is to rank the multiresponse observations, $\mathbf{y}_j (j = 1, \cdots, n)$, in terms of their Euclidean distance, $||\mathbf{y}_j - \mathbf{y}^*||$, (or equivalently, the squared Euclidean distance, $(\mathbf{y}_j - \mathbf{y}^*)'(\mathbf{y}_j - \mathbf{y}^*)$) from some robust estimate of location, $\mathbf{y}^*$. Next, a subset of the observations whose ranks are the smallest $100(1 - \alpha)\%$ are chosen and used for computing a sum-of-products matrix

$$\mathbf{A}_0 = \sum_{\substack{l \in \text{chosen subset} \\ \text{of observations}}} (\mathbf{y}_l - \mathbf{y}^*)(\mathbf{y}_l - \mathbf{y}^*)'. \tag{7}$$

(The fraction α of the observations not included in $\mathbf{A}_0$ is assumed to be small enough to ensure that $\mathbf{A}_0$ is non-singular.) Then all n observations are ranked in terms of the values of the quadratic form $(\mathbf{y}_j - \mathbf{y}^*)'\mathbf{A}_0^{-1}(\mathbf{y}_j - \mathbf{y}^*)$. Now a subset of the observations whose ranks are the smallest $100(1 - \beta)\%$ may be chosen and employed for defining a robust estimator of the covariance matrix

$$\mathbf{S}_3^* = \frac{k}{[n(1 - \beta)]} \sum_{\substack{r \in \text{chosen subset} \\ \text{of observations}}} (\mathbf{y}_r - \mathbf{y}^*)(\mathbf{y}_r - \mathbf{y}^*)', \tag{8}$$

where k is a constant that will hopefully make the estimator sufficiently unbiased, and again β has to be small enough so that $n(1 - \beta) > p$ and $\mathbf{S}_3^*$ is non-singular with probability 1. It may be convenient, but it is not imperative, to have $\alpha = \beta$. The above steps can be repeated using the sum-of-products on the right-hand side of (8) in place of $\mathbf{A}_0$, repeating the ranking of the observations, subselecting a major fraction of them for obtaining a further estimate, and iterating the process until a stable estimate is obtained. The limited experience of the authors with this method seems to suggest that unless α, β, and n are moderately large (viz. α and $\beta \geq 0.2$ and $n \geq 50$) and unless the underlying correlation structure for the observations is nearly singular, iterations will not be necessary to improve the estimate defined by (8).

The scheme involved in obtaining $\mathbf{S}_3^*$ depends on having an estimate $\mathbf{y}^*$ of location, and any of the location estimators discussed in section 2.1 may be utilized. Exactly as in the estimation of univariate variance, however, the location estimation can be circumvented by working with pairwise differences, $\mathbf{y}_i - \mathbf{y}_{i'}$, of the observations. Specifically, an estimator $\mathbf{S}_4^*$ can be obtained by repeating each of the steps involved in getting $\mathbf{S}_3^*$, with $\mathbf{y}_i - \mathbf{y}^*$ replaced by $\mathbf{y}_i - \mathbf{y}_{i'}$, working with rankings of these $\frac{1}{2}n(n-1)$ differences, and obtaining as an estimator analogous to $\mathbf{S}_3^*$ the matrix

$$\mathbf{S}_4^* = \frac{k'}{n(n-1)(1-\beta)} \sum_{\substack{r,s \text{ chosen subset} \\ \text{of observations}}} (\mathbf{y}_r - \mathbf{y}_s)(\mathbf{y}_r - \mathbf{y}_s)'. \tag{9}$$

Just as in the univariate variance situation discussed in section 2.2.1, this estimator may be poor when a large fraction of the observations are outliers.

The multiplicative constants k and k' in equations (8) and (9) are not as simply conceptualized or computed as the constants involved in the trimmed or Winsorized variances and covariances. An approach for estimating these constants via estimates of the scale parameter of the approximate gamma distribution of quadratic forms, such as $(\mathbf{y}_i - \mathbf{y}^*)'\mathbf{S}_3^{*-1}(\mathbf{y}_i - \mathbf{y}^*)$ and $(\mathbf{y}_i - \mathbf{y}_{i'})'\mathbf{S}_4^{*-1}(\mathbf{y}_i - \mathbf{y}_{i'})$, seems to be worth pursuing.

$\mathbf{S}_3^*$ and $\mathbf{S}_4^*$ both are estimates that involve a kind of trimming of the multivariate sample and the amount of trimming (viz. values of α and β) which is appropriate may be expected to depend on both n and p. More work is needed to develop bases for recommending reasonable values of α and β in practice.

Since the issue of discarding observations is likely to be particularly uneconomical in the multiresponse situation, where n may not often be much greater than p, the appeal of a 'Winsorizing-type' of scheme may be stronger. Specifically, rather than omitting any observation, one may wish to weight each observation reciprocally by some measure of how 'far' it is from the 'middle' of the data. Thus one could derive a sum-of-products matrix from the weighted observations:

$$\mathbf{A}_5^* = \sum_{i=1}^{n} w_i(\mathbf{y}_i - \mathbf{y}^*)(\mathbf{y}_i - \mathbf{y}^*)', \tag{10}$$

where $\mathbf{y}^*$ is a robust estimate of location and one possible choice for the weights is given by taking w_i as the reciprocal of $||\mathbf{y}_i - \mathbf{y}^*|| + \epsilon$, the ϵ being a small positive quantity to avoid difficulties when $||\mathbf{y}_i - \mathbf{y}^*||$ is negligibly small. The use of a weighted Euclidean metric, developed iteratively, in place of $||\mathbf{y}_i - \mathbf{y}^*||$ in the definition of w_i may be especially important when the variables have markedly different scales and high correlations. For estimating the covariance matrix, one might use $\mathbf{S}_5^* =$ a scalar multiple of $\mathbf{A}_5^*$.

An analogous approach in terms of pairwise differences of the observations that avoids the estimation of location is again possible. Explicitly, one could define a sum-of-products matrix $\mathbf{A}_6^*$ (and an associated estimator of the covariance matrix, $\mathbf{S}_6^*$) utilizing $\mathbf{y}_i - \mathbf{y}_{i'}$ in place of $\mathbf{y}_i - \mathbf{y}^*$ in (10) with the

obvious modifications in the definition of the weights, w_i , and the range of the summation sign in (10). Tukey [1971] has suggested more sophisticated versions of the weights, w_i , involving an iterative scheme for choosing final values of them. More work is needed on all of the issues involved in $\mathbf{S}_5^*$ and $\mathbf{S}_6^*$. Incidentally, these estimators are expected to be non-singular with probability 1 if $n \not< p$.

To give some feeling for how these estimators might work in practice, the results of calculating $\mathbf{S}_3^*$ using $\mathbf{y}_{HL}^*$ and $\alpha = \beta = 0.10$ from the $n = 20$ and $\rho = 0.5$ Monte Carlo data are presented here. At the end of one iteration of the procedure the estimated variances and covariance were $0.726k$, $0.756k$, and $0.319k$, while after the second iteration the values were $0.730k$, $0.757k$, and $0.323k$. Thus the correlation has improved slightly from 0.431 to 0.434 by repeating the procedure. The appropriate value of k, which may be a function of n, α, β, and ρ, appears in this example to be about 1.35. In more complex examples, where the scales of the variables may differ widely, several iterations may be necessary before a satisfactory single k value can be found.

3. ANALYSIS OF MULTIDIMENSIONAL RESIDUALS

Given some summarizing fit to a body of multiresponse data, there exists, in principle, a vector of multivariate residuals between the data and the fit; but, more than in the univariate case, there is the important issue of how to express these multivariate residuals. Though experience is still rudimentary on these matters, some things can be done and in this section the discussion will be concerned with some statistical methods for analyzing multivariate residuals.

For the discussion here and in section 4, it would be convenient to distinguish two broad categories of statistical analyses of multiresponse problems: (1) the analysis of internal structure and (2) the analysis of superimposed or extraneous structure. The first category includes techniques, such as principal components, factor analysis, and multidimensional scaling, which are useful for studying internal dependencies and for reduction of the dimensionality of response. Multivariate multiple regression and multivariate analysis of variance, which are the classical techniques for investigating and specifying dependence of multiresponse observations on design characteristics or extraneous independent variables, are examples belonging to the second category.

Each category of analysis gives rise to multivariate residuals. For instance, linear principal components analysis may be viewed as fitting a set of mutually orthogonal hyperplanes to the data so as to minimize, at each stage, the sum of squares of orthogonal deviations of the observations from each plane in turn (cf. Pearson [1901]). At any stage, therefore, one has residuals which are perpendicular deviations of data from the fitted hyperplane. On the other hand, in analyzing superimposed structure (i.e. the second category above) by multivariate multiple regression, one has the well-known least

squares residuals, viz. (observations) − (predictions from a least squares fit). For purposes of data analysis, it is often desirable to use the least squares residuals as input to a principal components analysis which, in turn, would lead to the orthogonal residuals mentioned earlier. Augmenting multivariate multiple regression fitting by a principal components transformation of the residuals from fit may help in describing statistical correlations in the errors of the combined original variables, or in indicating inadequacies in the fit of the response variables by the design variables.

The two subsections which follow are concerned, respectively, with principal components residuals and least squares residuals.

3.1. *Principal components residuals*

The starting point in analyzing internal structure by principal components is the set of n p-dimensional observations, the columns of the $p \times n$ matrix $\mathbf{Y}$, which are considered for purposes of the analysis as an unstructured multivariate sample. [Note: The observations may either be original data or derived data such as outputs of other analyses (e.g. least squares residuals from a multivariate analysis of variance).] The usual sample covariance matrix $\mathbf{S}$ or correlation matrix $\mathbf{R}$ may then be computed. In the case of $\mathbf{S}$, the linear principal components transformation of the data is given by

$$\mathbf{Z} = \mathbf{L}(\mathbf{Y} - \bar{\mathbf{Y}}), \tag{11}$$

where the p rows of the orthogonal matrix $\mathbf{L}$ are eigenvectors of $\mathbf{S}$ customarily chosen to correspond to its eigenvalues in descending order of magnitude, and $\bar{\mathbf{Y}} = \bar{\mathbf{y}} \cdot \mathbf{1}'$ is a $p \times n$ matrix all of whose columns are equal to the sample mean vector $\bar{\mathbf{y}}$. Each row, $\mathbf{l}_i'(i = 1, \cdots, p)$, of $\mathbf{L}$ provides a principal component coordinate and each row of $\mathbf{Z}$ gives the deviations of the projections of the original sample from the projection of the sample centroid $\bar{\mathbf{y}}$ onto a specific principal component coordinate. Using standardized variables as the starting point, the preceding definitions and descriptions would correspond to the principal components analysis of $\mathbf{R}$.

Viewed as a method of fitting linear subspaces, or as a statistical technique for detecting and describing possible linear singularities in the data, the interest would especially be in the projections of the data onto the principal component coordinates corresponding to the small eigenvalues (i.e. the last few rows of $\mathbf{Z}$). Thus, for instance, with $p = 2$ the essential concepts are illustrated in Figure A. In this figure, y_1 and y_2 denote the original coordinates and z_1 and z_2 denote the two principal components derived from the covariance matrix of the bivariate data. The straight line of closest fit to the data (where closeness is measured by the sum of squares of perpendicular deviations) is the z_1-axis. The orthogonal residual of a typical data point, P, as shown in the figure is the vector $\overrightarrow{QP}$ which is seen to be equivalent to the vector $\overrightarrow{O'P'}$, where P' is the projection of P onto the z_2-axis, the second principal component. More generally, with p-dimensional data, the projection

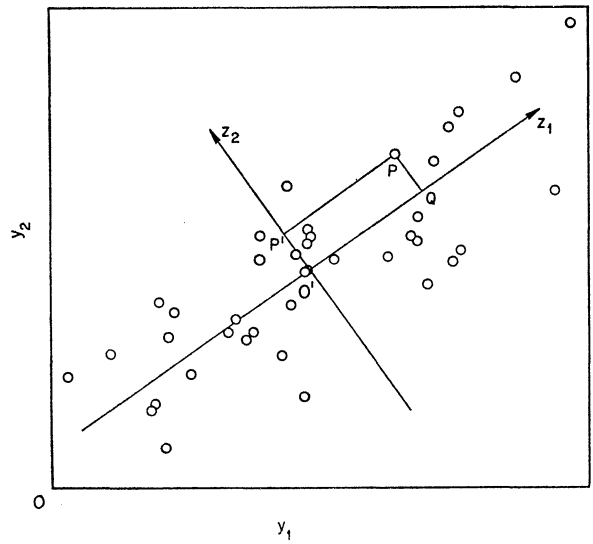

FIGURE A

PRINCIPAL COMPONENTS z_1 AND z_2 OF ORIGINAL VARIABLES y_1 AND y_2

onto the 'smallest' (i.e. the one with least variance) principal component would be relevant for studying the deviation of an observation from a hyperplane of closest fit while projections on the 'smallest' q principal component coordinates would be relevant for studying the deviation of an observation from a fitted linear subspace of dimensionality $p - q$.

For detecting lack of fit of individual observations, one method suggested by Rao [1964] is to study the sum of squared lengths of the projections of the observations on the last few (say q) principal component coordinates. For each initial observation $\mathbf{y}_j$ ($j = 1, \cdots, n$), the procedure consists of computing

$$d_j^2 = \sum_{i=p-q+1}^{p} [l_i'(\mathbf{y}_j - \bar{\mathbf{y}})]^2$$

$$= (\mathbf{y}_j - \bar{\mathbf{y}})'(\mathbf{y}_j - \bar{\mathbf{y}}) - \sum_{i=1}^{p-q} [l_i'(\mathbf{y}_j - \bar{\mathbf{y}})]^2, \tag{12}$$

and considering inappropriately large values of d_j^2 as indicative of a poor $(p - q)$-dimensional fit to the observation (or, equivalently, that the observation is possibly an aberrant one). An informal graphical technique, which may have value as a tool for exposing other peculiarities of the data besides assessing the fit, would be to make a gamma probability plot (cf. Wilk *et al.* [1962a]) of the d_j^2 using an appropriately chosen or estimated shape parameter. One method of obtaining a suitable estimate of the shape parameter would be to base it on a collection of the smallest observed d_j^2 as suggested in a different context by Wilk and Gnanadesikan [1964].

In addition to looking at a single summary statistic, such as d_j^2 above,

it may often be useful to study the projections of the data on the last few principal component coordinates (i.e. the last few rows of **Z** in (11)) in other ways. These might include: (i) Two- and three-dimensional scatter plots of bivariate and trivariate subsets of the last few rows of **Z** with points labeled in various ways such as by time if it is a factor. (ii) Probability plots of the values within each of the last few rows of **Z**. Because of the linearity of the transformation involved, it may not be unreasonable to expect these values to be more nearly normally distributed than the original data and normal probability plotting would provide a reasonable starting point for the analysis. This analysis may help in pinpointing specific 'smallest' principal component coordinates, if any, on which the projection of an observation may look abnormal and, thus, augment the earlier-mentioned gamma probability plotting analysis of the d_i^2 . (iii) Plots of the values in each of the last few rows of **Z** against certain distances in the space of the first few principal components. If, for example, most of the variability of a set of five-dimensional data is associated with the first two principal components, then it may be informative to plot the projections on each of the three remaining principal component axes against the distance from the centroid of each of the projected points in the two-dimensional plane associated with the two largest eigenvalues. This may show a certain kind of multidimensional inadequacy of fit; namely, if the magnitude of the residuals in the coordinates associated with the smaller eigenvalues is related to the clustering of the points in the two-dimensional space of the two eigenvectors corresponding to the largest two eigenvalues.

The above discussion of principal components residuals has been in terms of the usual linear principal components analysis. Similar concepts and approaches need to be developed for generalized principal components analysis, a technique proposed by Gnanadesikan and Wilk [1968; 1969] for detecting and describing possibly nonlinear relationships among the responses. With nonlinear functions involved, some care is required in defining, computing, and finally expressing statistically the perpendicular deviations of the observations from the fitted function.

A somewhat different but important issue is one that concerns the robustness of the suggested analyses. Clearly if an aberrant observation is detected then one may want to exclude it from the initial estimate of **S** (or **R**) and then repeat the process of obtaining and analyzing the principal components residuals. In some circumstances, one may also decide to use a robust estimate of the covariance (or correlation) matrix, such as the ones considered in section 2, even for the initial analysis, with the hope that the aberrant observations would become even more conspicuous in the subsequent analysis of residuals.

Example 1. To illustrate the use of some of the methods for analyzing principal components residuals, two sets of data are taken from a study (cf. Chen *et al.* [1970]) which was concerned with bases for groupings of corporations. As a part of the study, the appropriateness of prespecified groupings (e.g.

chemicals, oils, drugs, etc.) was examined initially, and a preliminary attempt was made to develop core groups of companies from an internal analysis of each prespecified category. One approach for forming core groups was to identify and eliminate outliers by studying the principal components residuals.

There were 14 variables per company per year in the study and Figure 1a shows a scatter plot of all the drug companies in the space of the last two principal components of the 14 $\times$ 14 correlation matrix derived from the observations on all 20 companies for the year 1963. Companies 9 and 8 are indicated as possible outliers with respect to the configuration of the remaining companies in this plot. Company 9 appears to be an outlier with respect to the last principal component in particular while company 8 seems to be a moderate outlier on the penultimate principal component.

The second set of data is from 23 drug companies for the year 1967 and is employed to illustrate the use of probability plotting of the elements in each of the last few rows of **Z**. Figure 1b shows a normal probability plot of the 10th sample principal component of the correlation matrix. The points corresponding to companies 11 and 19 are seen to be 'too large', deviating at the top right-hand end of the plot from the reasonably good linear configuration of the remaining points. The original data in this example exhibited considerable non-normality, and the earlier-mentioned aspect of improved normality induced by the principal components transformation is seen in Figure 1b by the linearity of the configuration of most of the points with just a mild indication of a distribution with shorter tails than the normal.

3.2. *Least squares residuals*

The usual descriptions of multivariate multiple regression (see, e.g., Roy *et al.* [1971]) involve the separate regressions of each of the multivariate responses on a common design or regressor matrix, yielding a matrix of estimated regression coefficients having certain joint statistical properties. Formally, the so-called multivariate general linear model may be stated as

$$\mathbf{Y}'_{n\times p} = \mathbf{X}_{n\times m}\boldsymbol{\Theta}_{m\times p} + \boldsymbol{\varepsilon}_{n\times p}\,, \tag{13}$$

where (i) the n rows of $\mathbf{Y}'$ are the multiresponse observations, (ii) the n rows of $\mathbf{X}$ are the n values of m independent or design variables, (iii) the elements of $\boldsymbol{\Theta}$ are the unknown regression coefficients or effects, and (iv) the rows of $\boldsymbol{\varepsilon}$ are uncorrelated p-dimensional random variables with certain assumed joint statistical distributional properties, e.g., mean zero and common covariance matrix.

The multivariate least squares residuals (to be hereafter referred to simply as residuals) are the n p-dimensional rows of

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y}' - \mathbf{X}\hat{\boldsymbol{\Theta}}, \tag{14}$$

where, according to the usual treatment of the subject, the matrix, $\hat{\boldsymbol{\Theta}}$, of estimated coefficients is obtained by taking for each of its columns the usual

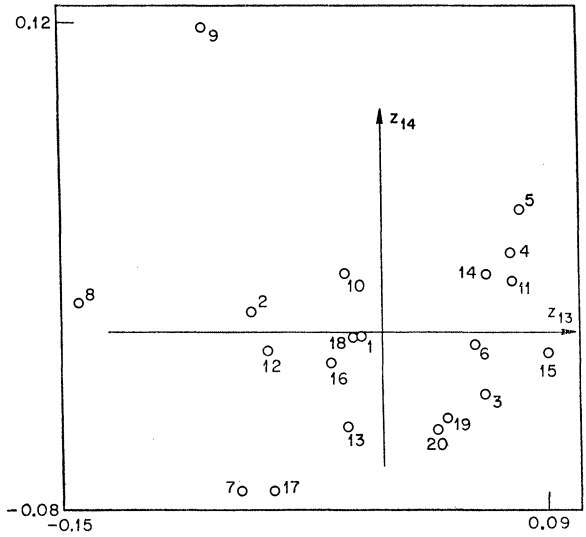

FIGURE 1a

LAST TWO PRINCIPAL COMPONENTS BASED ON THE CORRELATION MATRIX FOR 20 DRUG COMPANIES

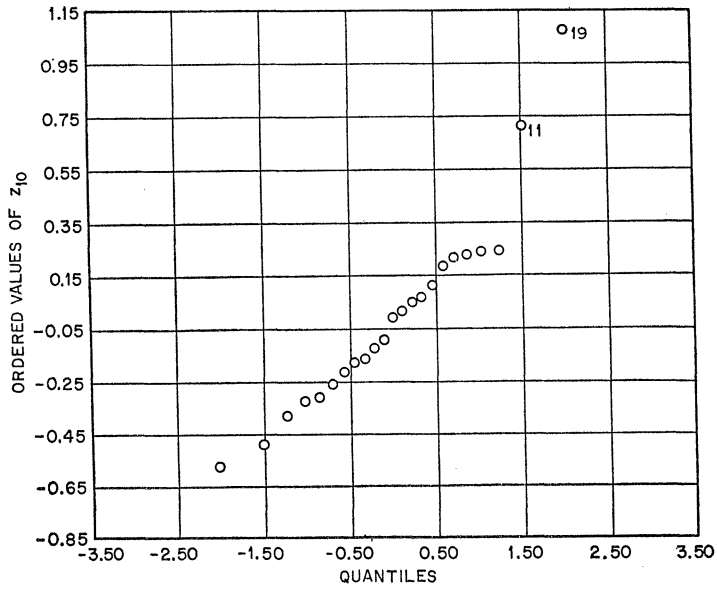

FIGURE 1b

NORMAL PROBABILITY PLOT OF THE TENTH PRINCIPAL COMPONENT BASED ON THE CORRELATION MATRIX FOR 23 DRUG COMPANIES

least squares estimates derived from analyzing each response separately. Depending on the structure of $\mathbf{X}$, there will be certain singularities among the residuals in that certain linear combinations of the rows of $\hat{\boldsymbol{\varepsilon}}$ will be zero. Depending on the correlational structure and functional dependencies among the p responses there could be singularities in the other direction (viz. the columns of $\hat{\boldsymbol{\varepsilon}}$), and an investigation of the existence and nature of such singularities may be pursued by principal components transformations (cf. section 3.1) of the p-dimensional residuals.

In some applications, there may be a natural ordering among the responses which might lead one to consider the use of a step-down analysis (cf. section 4.c of Chapter IV in Roy *et al.* [1971]). The analysis at each stage is a univariate analysis of a single response utilizing all the responses which have been analyzed at the preceding stages as covariates. At each stage, therefore, step-down residuals may be obtained from this approach and studied by any of the available techniques for analyzing univariate least squares residuals.

Recently, Larsen and McCleary [1970] have proposed the concept of partial residuals and ways of using them. Entirely analogous definitions of multivariate partial residuals, and methods of analyzing them, may be suggested.

As a first cut at analyzing the residuals defined in (14), one may wish to consider the entire collection of them as an unstructured multivariate sample. Sometimes such a view may be more appropriate for subsets of the residuals than for the totality of them. For instance, in a two-way table the residuals within a particular row (or column) may be considered as an unstructured sample. At any rate, with such a view, one can then employ methods applicable to the study of unstructured multivariate samples, including: (i) Separate plotting of uniresponse residuals, perhaps against values of certain independent or extraneous variables (e.g. time), or against the predicted values. (ii) Two-dimensional scatter plotting of bivariate subsets of the residuals and three-dimensional stereoscopic representations of trivariate subsets. These would be reasonably good graphical tools for studying cohesiveness, groupings, possible outliers, and general shape. (iii) One-dimensional probability plotting of the uniresponse residuals. Full-normal plots of the uniresponse residuals or half-normal plots of their absloute values (or, equivalently, χ_1^2 plots of squared residuals) provide natural starting points. Residuals generally seem to tend to be 'supernormal' or at least more normally distributed than original data and such probability plots may be useful in delineating outliers or other peculiarities in the data. (iv) The use of one, or preferably several, distance functions to convert the multiresponse residuals to single numbers, followed by the probability plotting of these. A common, useful class of distance functions is that of positive semi-definite quadratic forms, $\mathbf{e}'\mathbf{M}\mathbf{e}$, where $\mathbf{e}$ is a p-dimensional residual and the matrix, $\mathbf{M}$, may itself be derived from the collection of residuals. Methodology for developing a gamma probability plot, with a suitably estimated shape parameter, for the values of such a quadratic

form has been suggested by Wilk and Gnanadesikan [1964]. Their work was concerned with a procedure for intercomparing the relative 'sizes' of single-degree-of-freedom contrast vectors and, in the present context, the p-dimensional residuals are being considered as approximately single-degree-of-freedom quantities. The procedure would consist of obtaining and ordering the n values of the quadratic form for a specific choice of $\mathbf{M}$, determining the maximum likelihood (ML) estimate of the shape parameter of the null gamma distribution in question from a set of the smallest of the ordered values (cf. Wilk *et al.* [1962b]) and plotting the ordered values against the corresponding quantiles of the gamma distribution with the estimated shape parameter (cf. Wilk *et al.* [1962a]). Under null conditions (i.e. the absence of any peculiarities and the conformity to all assumptions), the configuration of such a plot would be expected to be linear. Deviations from null conditions, such as the presence of an outlier or heteroscedasticity, would be indicated by departures of the configuration from linearity. For instance, an aberrant observation may be expected to yield a residual for which the associated quadratic form value would be unduly large, thus leading to a departure of the corresponding point from the linearity of the other points on the gamma probability plot (cf. Example 3 below). Heteroscedasticity would be indicated by a configuration which is piecewise linear, with the points corresponding to the residuals derived from observations with the same covariance structure belonging to the same linear piece.

Example 2. The example, taken from Gnanadesikan and Wilk [1969], illustrates the utility of gamma probability plotting of quadratic forms in the multivariate resideuals for the simplest case of an unstructured sample. The residuals here would be just deviations of the individual multiresponse observations from the sample mean vector, $\mathbf{e}_j = \mathbf{y}_j - \bar{\mathbf{y}}$ $(j = 1, \cdots, n)$, a special case which has been considered by Cox [1968] and Healy [1968].

The computer generated trivariate data (cf. Example 5 of Gnanadesikan and Wilk [1969]) was obtained by perturbing (viz. adding normal errors) the points on the surface of a paraboloid. Figures 2a–c show the three scatter plots of the data for each possible pair of the three coordinates. None of these reveal the cup-shaped nature of the data configuration. Figure 2d shows a gamma probability plot of the n values, $\mathbf{e}_j'\mathbf{S}^{-1}\mathbf{e}_j$ $(j = 1, \cdots, n)$, where $\mathbf{S}$ is the sample covariance matrix. The shape parameter used for obtaining this plot was specified as $\eta = \frac{3}{2}$, using the intuitive argument that the quadratic form, $(\mathbf{y}_j - \bar{\mathbf{y}})'\mathbf{S}^{-1}(\mathbf{y}_j - \bar{\mathbf{y}})$, would be distributed approximately as a χ^2 with p $(= 3$ in this example) D.F. (cf. also the discussion in section 4.1 below). The existence, though not the detailed nature, of a 'hole' in the cup-shaped data configuration is exhibited in Figure 2d by the nonzero intercept of the plot, reflecting the lack of near-zero values of the quadratic form.

Example 3. The data derives from an experiment on long-term aging of a transistor device used in submarine cable repeaters (cf. Abrahamson *et al.* [1969]). Sets of 100 devices, in a configuration of 10 rows by 10 columns,

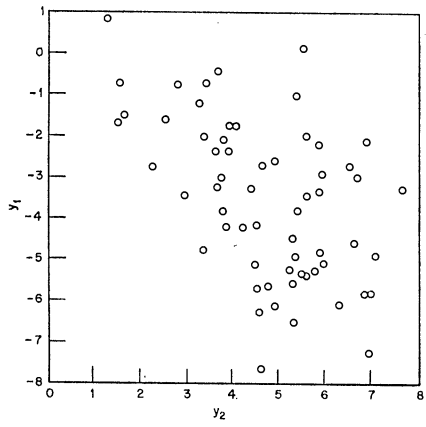

FIGURE 2a

SCATTER PLOT OF y_1 VS. y_2 FOR CUP-SHAPED DATA

were aged and a characteristic called the gain of each device was obtained at each of several test periods. An initial transformation to logarithms was made and the aging phenomenon of interest was then the behavior of the log-gain as a function of time. One approach to studying the aging behavior for purposes of identifying devices with peculiar aging characteristics was to fit a polynomial (specifically, a cubic was used) to the data on log-gain versus time for each device and to study the fitted coefficients by analysis of variance techniques. Separate univariate analyses of variance of each coefficient, as well as a multivariate analysis of variance of the four coefficients simultaneously, were performed. The multivariate approach was employed partly because of the high intercorrelations that were observed among the fitted coefficients. It was not used as a substitute for the separate

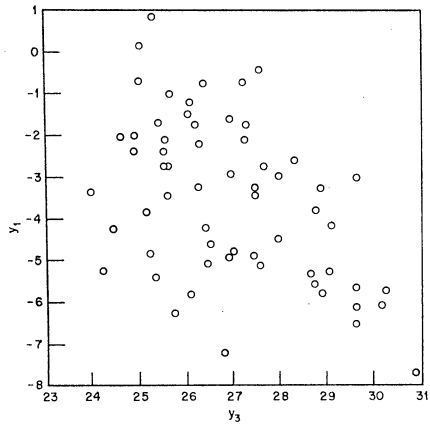

FIGURE 2b

SCATTER PLOT OF y_1 VS. y_3 FOR CUP-SHAPED DATA

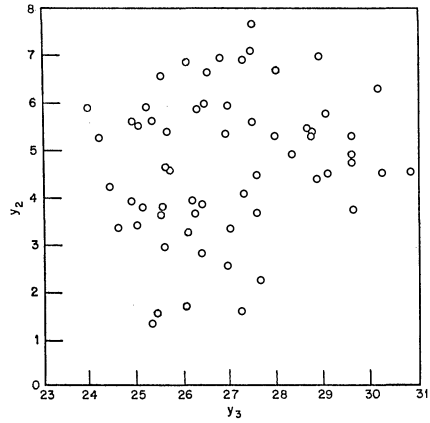

FIGURE 2c

SCATTER PLOT OF y_2 VS. y_3 FOR CUP-SHAPED DATA

univariate analyses of the individual coefficients. For present purposes, attention is confined to the multivariate approach.

A simple one-way (viz. rows and columns-within-rows were the sources of variation) multivariate analysis of variance, when used as a means for obtaining formal tests of hypotheses, revealed very little. None of the usual tests (cf. Chapter IV of Roy *et al.* [1971]) of the hypothesis of no row effects had an associated p-value smaller than 0.3. The danger in basing an analysis solely on such tests which are based on single summary statistics is brought out by the use of the informal techniques described earlier in this section.

Figure 3a shows a gamma probability plot of the 100 values of a quadratic form, $\mathbf{e}_j'\mathbf{S}^{*-1}\mathbf{e}_j$ ($j = 1, \cdots, 100$), in the four-dimensional residuals. The covariance matrix, $\mathbf{S}^*$, of the residuals is a robust estimate (of the type discussed in section 2.2.3) obtained from the residuals themselves. [Note:

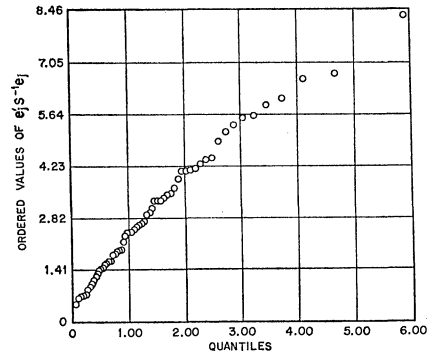

FIGURE 2d

GAMMA PROBABILITY PLOT OF GENERALIZED SQUARED DISTANCES DERIVED FROM CUP-SHAPED DATA (SHAPE PARAMETER = 1.5)

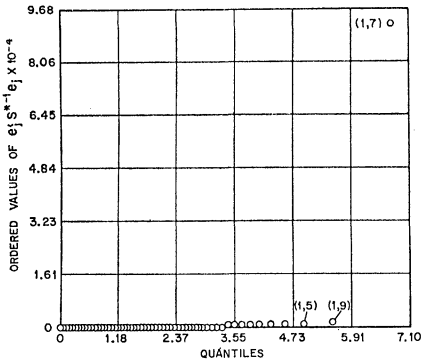

FIGURE 3a

GAMMA PROBABILITY PLOT OF QUADRATIC FORM VALUES FOR 100 RESIDUALS (EST. SHAPE PARAMETER = 1.663)

Since $\mathbf{S}^{*-1}$ is common to all 100 values of the quadratic form being analyzed, it is not necessary to multiply $\mathbf{S}^*$ by the unbiasing constant for the present application.] The shape parameter required for the plot was estimated by maximum likelihood based on the 50 smallest values of the quadratic form considered as the 50 smallest order statistics in a random sample of size 100.

The point which stands out clearly from the configuration of the others in Figure 3a corresponds to the 7th device in the first row and the implication is that the four-dimensional residual for this device is inordinately 'large', i.e. a possibly aberrant observation has been pinpointed! This residual (and other such if they exist) has, of course, contributed to the estimate of the columns-within-rows dispersion matrix which was employed as the error dispersion matrix in the formal tests of significance mentioned earlier. The effect would be to inflate the error dispersion inappropriately and it is, therefore, not surprising that the tests revealed no significant departures from the null hypothesis. Upon verification, the aging configuration of device 7 in row 1 was found to be indeed abnormal relative to the behavior of the majority of devices.

To facilitate further study of the residuals, a replot, shown in Figure 3b, may be made of the 99 points left after omitting the point corresponding to the aberrant device. The configuration on this plot might lead one to conclude that device 9 (the one from which the top right-hand corner point derives) and also the other devices (1–6, 8, and 10) in row 1 are suspect, i.e. all 10 devices in row 1 are associated with peculiar residuals. Such a conclusion, however, may not be warranted and the discussion that follows will clarify the issue involved. The analysis of the data is then continued in Example 4.

When the matrix $\mathbf{X}$ in (13) corresponds to more structured situations (e.g. a multiway classification), then there are at least two sources of statistical difficulties in analyzing the residuals. First, there are constraints on subsets of the residuals (e.g. the sum of the residuals in a row of a two-way table is the null vector) which imply correlations among the residuals. Second, the

presence of outliers might seriously bias the usual effects which are subtracted from an observation (e.g. row, column, and overall mean vectors in a two-way classification) so as to mask the local effect of an outlier on the corresponding residual. The first source of difficulty (viz. the singularities among residuals) is especially critical when the numbers of levels of the factors involved (e.g. the number of rows or columns in a two-way table) are small, but the second source could be important even when the factors each have a moderate number of levels.

Thus in Example 3 above, the extreme outlier (viz. the observation for device 7 in row 1) may have so badly biased the mean vector for the first row that all the residuals (= (observation vector) − (row mean vector)) in that row have been unduly biased. If the outlier is extreme enough this can indeed happen, and a method is needed for insuring against such masking effects of the outliers on the residuals.

One way of accomplishing this is to combine the ideas and methods of robust estimation with the desirability of analyzing the residuals, i.e. combine summarization and exposure. Specifically, instead of using the usual least squares estimates of the elements of Θ in the linear model (13), one could use robust estimates of them thus obtaining $\hat{\Theta}^*$, and then define a set of *modified residuals* as the rows of

$$\hat{\varepsilon}^* = Y' - X\hat{\Theta}^*. \tag{15}$$

For the usual analysis of variance situations, since the rows of Θ in (13) are multivariate location type parameters, the rows of $\hat{\Theta}^*$ may be estimated by using robust estimators of location such as those discussed in section 2.1. In fact, since every robust multivariate location estimator with the exception of the one due to Gentleman [1965] is based on using robust univariate location estimators for each response, essentially $\hat{\Theta}^*$ is a matrix each of whose elements, $\hat{\theta}^*_{ij}$, is a robust estimator of a univariate location type parameter.

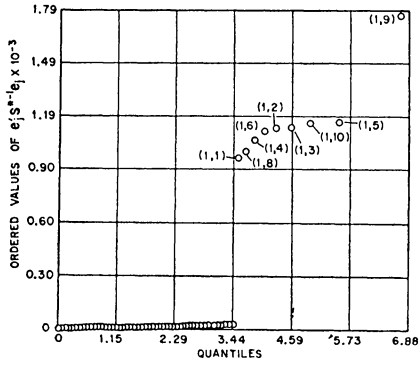

FIGURE 3b

REPLOT OF QUADRATIC FORM VALUES OMITTING POINT (1,7)

Example 4. To illustrate the use of modified residuals, the data used in Example 3 is employed again. Instead of using the row mean vectors for defining the residuals, the vector of midmeans, $\mathbf{y}^*_{T(.25)}$, discussed in section 2.1, for each row is used and the modified four-dimensional residuals are obtained as the difference between the four-dimensional observation (viz. the four coefficients of the aging curve for a device) and the vector of midmeans for the row in which the observation appears.

The 100 four-dimensional modified residuals thus obtained in this example can then be analyzed by the gamma probability plotting technique described and illustrated earlier in the context of analyzing the regular residuals. Figure 4a shows a gamma probability plot of the 100 values of a quadratic form in the modified residuals, $\mathbf{e}_j^{*\prime}\mathbf{S}^{*-1}\mathbf{e}_j^*$ ($j = 1, \cdots, 100$), where $\mathbf{S}^*$ as before is a robust estimate of the covariance matrix, and the shape parameter required for the plot is estimated once again using the smallest 50 observed values of the quadratic form. In Figure 4a, the point corresponding to device 7 in row 1 again stands out and Figure 4b shows a replot obtained after omitting this point. By comparing Figure 4b with Figure 3b it is seen that the biasing effect on all the residuals in the first row caused by the extremely deviant observation for device 7 in that row is no longer evident. The configuration in Figure 4b may be used to delineate additional outliers, such as device 1 in row 7, by looking for points in the top right-hand corner that deviate noticeably from the linear configuration of the points in the lower left-hand portion of the picture.

The modified residuals defined by (15) will not necessarily satisfy the constraints satisfied by the usual residuals. For example, in a two-way classification, they will not necessarily add up to the null vector either by rows, or by columns, or even across all cells. The modified residuals do not form a cohesive group unless there are no outliers in the data and, in the latter case, the usual least squares estimator $\hat{\Theta}$ and the robust estimator $\hat{\Theta}^*$ would not be very different, so that the usual residuals, $\hat{\varepsilon}$, and the modified

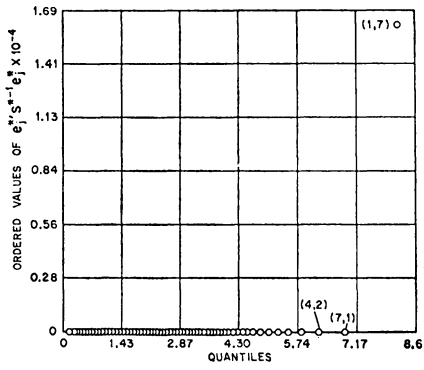

FIGURE 4a

GAMMA PROBABILITY PLOT OF QUADRATIC FORM VALUES FOR 100 MODIFIED RESIDUALS (EST. SHAPE PARAMETER = 2.409)

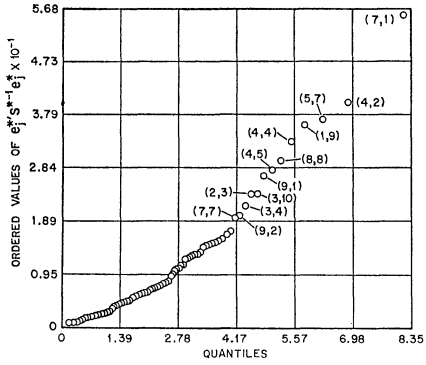

FIGURE 4b

REPLOT OF QUADRATIC FORM VALUES OMITTING POINT (1,7)

residuals, $\hat{\varepsilon}^*$, would also be expected to be very similar when there are no outliers. The main use of the modified residuals is, in fact, to accentuate the presence of outliers and, as such, the fact that they do not satisfy the same constraints as the usual residuals is perhaps unimportant. If, however, one desires to have modified residuals satisfy these constraints as nearly as possible, then iterating the analysis in certain ways might help. Tukey [1970] has suggested such a scheme for using midmeans in analyzing multiway tables with uniresponse data, and an extension of this approach to the multiresponse case seems feasible.

4. THE DETECTION OF MULTIVARIATE OUTLIERS

In the previous section, ways of pinpointing maverick observations through an analysis of multivariate residuals were discussed. In this section some additional techniques for detecting multivariate outliers will be put forth.

The consequences of having defective responses in a multivariate sample are intrinsically more complex than in the much-discussed univariate case. (see, e.g., David [1970] and references therein.) One reason for this is that a multivariate outlier can distort not only measures of location and scale but also those of orientation (i.e. correlation). A second reason is that it is much more difficult to characterize the multivariate outlier. A single univariate outlier may be typically thought of as 'the one which sticks out on the end', but no such simple idea suffices in higher dimensions. A third reason is the variety of types of multivariate outliers which may arise: a vector response may be faulty because of a gross error in one of its components or because of systematic mild errors in all of its components.

The complexity of the multivariate case suggests that it would be fruitless to search for a truly omnibus outlier protection procedure. A more reasonable approach seems to be to tailor detection procedures to protect against specific types of situations, e.g., correlation distortion, thus building up an arsenal

of techniques with different sensitivities. This approach recognizes that an outlier for one purpose may not necessarily be one for another purpose! However, if several analyses are to be performed on the same sample, the result of selective segregation of outliers should be a more efficient and effective use of the available data.

It is essential that the procedures be computationally inexpensive enough to allow for routine screening of large data sets. Those which can simultaneously expose other features of the data, such as distributional peculiarities, have added economic appeal.

Following the dichotomy of multivariate methods mentioned in section 3, the proposed procedures will be presented under the general headings of internal and external analysis techniques. In the former category are those techniques, such as principal components analysis, which are appropriate for examining an unstructured sample of data; in the latter category are techniques, like canonical correlation analysis, which are applicable in the presence of some superimposed structure.

It should be apparent that all of the procedures are open to refinement through iteration or robustification.

4.1. *Internal analysis techniques*

A fundamental way of displaying multivariate data and in particular of uncovering outliers is through two- and three-dimensional plots of the original and the principal component variables.

Of the principal components, the first and last few are usually the most interesting. The first ones are especially sensitive to outliers which are in-

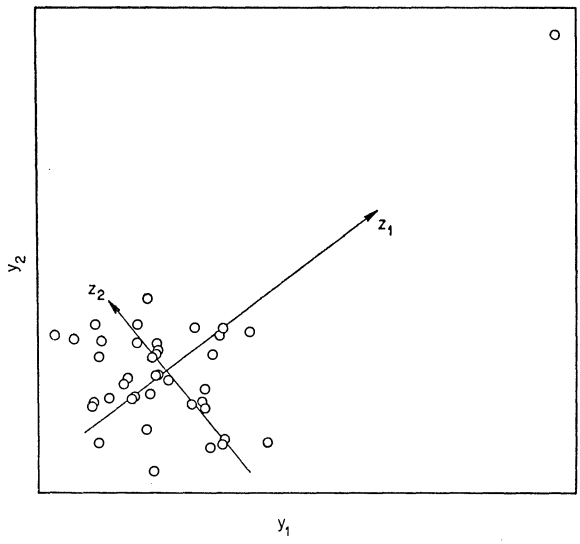

FIGURE B

EXAMPLE OF AN OUTLIER INFLATING VARIANCES AND COVARIANCES

appropriately inflating variances and covariances (if working with **S**) or correlations (if working with **R**). A pictorial representation of such an outlier is shown in Figure B. Motivation for looking at the last few principal components has been given in section 3.1. The kind of wild value which can be detected along these axes is one which is adding insignificant dimensions or obscuring singularities in the data. An example of this type of outlier is shown near the bottom of Figure C.

Probability plots and standard univariate outlier tests can be made on the rows of either **Y** or **Z**. The accumulated evidence may provide clues concerning such characteristics as the 'dimensionality' of any outliers discovered.

Also useful are easily calculated univariate statistics which measure the contribution of the individual observations (or combinations of the observations) to specific multivariate effects. The generated values can be studied graphically, often with a probability plot, in spite of minor difficulties due to correlations among the statistics. Some examples of such statistics are given below. (The subscript '$-j$' means that the jth observation has been deleted, and $\mathbf{A} = (n-1)\mathbf{S}$.)

$$r_{-j} = \frac{(n-1)a_{12} - n(y_{1j} - \bar{y}_1)(y_{2j} - \bar{y}_2)}{[(n-1)a_{11} - n(y_{1j} - \bar{y}_1)^2]^{\frac{1}{2}}[(n-1)a_{22} - n(y_{2j} - \bar{y}_2)^2]^{\frac{1}{2}}} \qquad j = 1, \cdots, n, \tag{16}$$

the correlation coefficients for variables 1 and 2 based on $n-1$ observations, for detecting outliers which are distorting the estimated correlation.

$$q_j^2 = \left(\frac{n}{n-1}\right)[\text{tr}\,(\mathbf{A}) - \text{tr}\,(\mathbf{A}_{-j})] = (\mathbf{y}_j - \bar{\mathbf{y}})'(\mathbf{y}_j - \bar{\mathbf{y}}) \quad j = 1, \cdots, n, \tag{17}$$

or the squared Euclidean distances of the $\mathbf{y}_j$'s from $\bar{\mathbf{y}}$, for isolating observations which are excessively inflating the overall scale. (Here tr stands for the trace.)

$$t_j^2 = \sum_i c_i[\mathbf{l}_i'(\mathbf{y}_j - \bar{\mathbf{y}})]^2 = (\mathbf{y}_j - \bar{\mathbf{y}})'\mathbf{S}(\mathbf{y}_j - \bar{\mathbf{y}}) \qquad j = 1, \cdots, n, \tag{18}$$

where c_i is the eigenvalue of **S** corresponding to $\mathbf{l}_i$. t_j^2 is a weighted sum of squares of the elements of the jth column of **Z**. Together they have the property that $\sum t_j^2 = (n-1)\sum c_i^2$, in contrast with the q_j^2 which sum to $(n-1)\sum c_i$. The t_j^2's are useful for determining which observations have the greatest influence on the orientation and scale of the first few principal components of **S**.

$$u_j^2 = \sum_i c_i\left[\frac{\mathbf{l}_i'(\mathbf{y}_j - \bar{\mathbf{y}})}{||\mathbf{y}_j - \bar{\mathbf{y}}||}\right]^2 = \frac{(\mathbf{y}_j - \bar{\mathbf{y}})'\mathbf{S}(\mathbf{y}_j - \bar{\mathbf{y}})}{(\mathbf{y}_j - \bar{\mathbf{y}})'(\mathbf{y}_j - \bar{\mathbf{y}})} \qquad j = 1, \cdots, n, \tag{19}$$

a collection of weighted sums of squares of the cosines of the angles between the eigenvectors of **S** and the centered data. u_j^2 is similar in spirit to t_j^2 except that the emphasis is more on orientation and less on scale.

$$v_j^2 = \sum_i c_i^{-1}\left[\frac{\mathbf{l}_i'(\mathbf{y}_j - \bar{\mathbf{y}})}{||\mathbf{y}_j - \bar{\mathbf{y}}||}\right]^2 = \frac{(\mathbf{y}_j - \bar{\mathbf{y}})'\mathbf{S}^{-1}(\mathbf{y}_j - \bar{\mathbf{y}})}{(\mathbf{y}_j - \bar{\mathbf{y}})'(\mathbf{y}_j - \bar{\mathbf{y}})} \qquad j = 1, \cdots, n, \tag{20}$$

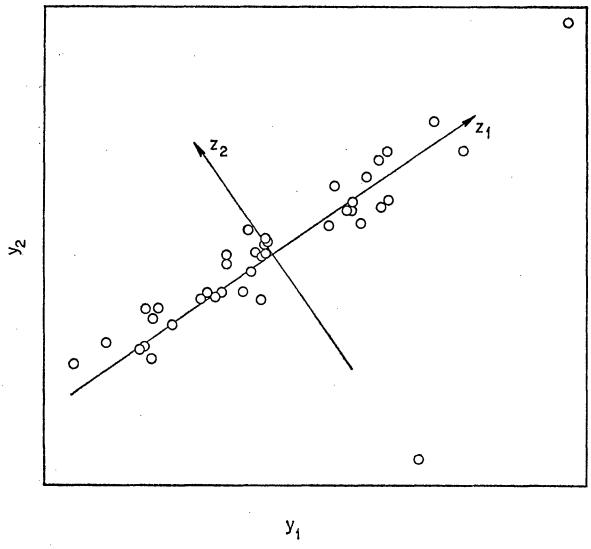

FIGURE C

EXAMPLE OF AN OUTLIER OBSCURING A NEAR SINGULARITY

which are like the u_i^2 but with the inverse weighting. This statistic measures the relative contributions of the observations on the orientations of the last few principal components.

$$d_{i0}^2 = \sum_i c_i^{-1}[l_i'(\mathbf{y}_i - \bar{\mathbf{y}})]^2 = (\mathbf{y}_i - \bar{\mathbf{y}})'\mathbf{S}^{-1}(\mathbf{y}_i - \bar{\mathbf{y}}) \qquad j = 1, \cdots, n, \qquad (21)$$

the generalized squared distances of the $\mathbf{y}_i$ from $\bar{\mathbf{y}}$ for uncovering observations which lie far afield from the general scatter of points.

$$d_{ij'}^2 = \sum_i c_i^{-1}[l_i'(\mathbf{y}_i - \mathbf{y}_{j'})]^2$$

$$= (\mathbf{y}_i - \mathbf{y}_{j'})'\mathbf{S}^{-1}(\mathbf{y}_i - \mathbf{y}_{j'}) \qquad j < j' = 1, \cdots, n, \qquad (22)$$

the generalized squared interpoint distances. These contain the same information as the d_{j0}^2 plus more detail.

q_i^2, t_i^2, and d_{i0}^2 are members of a general class of quadratic forms $(\mathbf{y}_i - \bar{\mathbf{y}})'\mathbf{S}^b(\mathbf{y}_i - \bar{\mathbf{y}})$. As b increases above $+1$, more and more emphasis is being placed on the first few principal components. As b decreases below -1, the opposite stress is achieved. Similar comments may be made about u_i^2 and v_i^2 with respect to a general class of quadratic form ratios $(\mathbf{y}_i - \bar{\mathbf{y}})'\mathbf{S}^b(\mathbf{y}_i - \bar{\mathbf{y}})/(\mathbf{y}_i - \bar{\mathbf{y}})'(\mathbf{y}_i - \bar{\mathbf{y}})$ and values of b above and below zero. Additional flexibility with either of these classes can be achieved by replacing $\mathbf{y}_i - \bar{\mathbf{y}}$ with $\mathbf{y}_i - \mathbf{y}_{j'}$.

Statistics similar to t_i^2 , u_i^2 , and v_i^2 can be constructed when $\mathbf{R}$ is used instead of $\mathbf{S}$. d_{i0}^2 and $d_{ij'}^2$ require no such adjustment as they are invariant under non-singular transformations of $\mathbf{Y}$.

If Fisher's transformation, $\frac{1}{2}\log[(1 + r_{-i})/(1 - r_{-i})]$, is applied to the

r_{-i}, then for bivariate normal data with no outliers one would expect a roughly linear normal probability plot of the transformed values. The influence of a single outlier would be to inflate (or deflate) unduly all but one of the r_{-i}, and one hopes that the plot will reveal this effect.

Generally speaking, gamma-type probability plots with estimated shape parameters are reasonable starting points for the statistics q_i^2, t_i^2, d_{i0}^2, and $d_{ii'}^2$. The u_i^2 and v_i^2 are more difficult to treat, but beta- or F-type probability plots may be worth trying. Cox [1968] and Healy [1968] have suggested a χ_p^2 probability plot for the d_{i0}^2 when the observations are approximately p-variate normal. An alternative (but closely related) approach for the study of the d_{i0}^2 is given below.

The exact marginal distributions of d_{i0}^2 and $d_{ii'}^2$ derived from multivariate normal samples are quite simple:

$$\frac{n}{(n-1)^2}\, d_{i0}^2 \quad \text{and} \quad \frac{1}{2(n-1)}\, d_{ii'}^2,$$

are beta variables with parameters $\frac{1}{2}p$ and $\frac{1}{2}(n-p-1)$. The proof is based on a theorem in Wilks ([1962] p. 562). Note that there are two levels of correlation among the $d_{ii'}^2$, one between distances with a single common subscript (asymptotically, it is $\frac{1}{2}$) and the other between distances with no common subscript (asymptotically, it is zero). Since the interest here is primarily in the behavior in the tail areas, it is preferable to transform the distances into F variables for probability plotting purposes (cf. Gnanadesikan *et al.* [1967]).

Wilks [1963] proposed that a test for a single outlier be based on

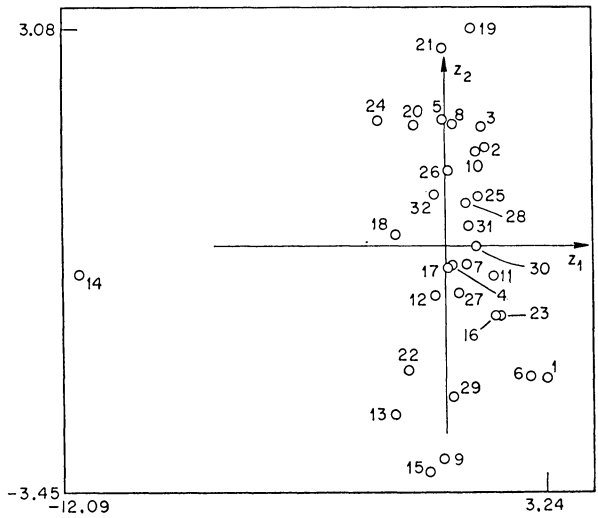

FIGURE 5

FIRST TWO PRINCIPAL COMPONENTS BASED ON THE CORRELATION MATRIX FOR 32 CHEMICAL COMPANIES

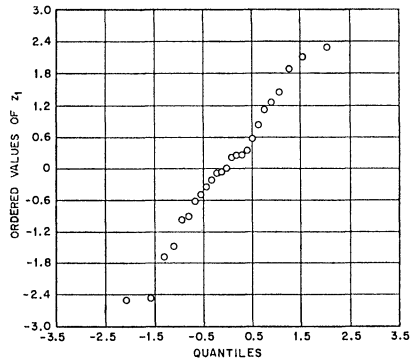

FIGURE 6a

NORMAL PROBABILITY PLOT OF FIRST PRINCIPAL COMPONENT IN EX. 6

$\max_i (|\bar{\mathbf{A}}_{-i}|/|\mathbf{A}|)$ which is equivalent to $\max_i d_{i0}^2$. The asymptotic distribution of this maximum, as well as $\max_{i,i'} d_{ii'}^2$, in multivariate normal samples, has been investigated by Siotani [1959].

Hierarchical clustering procedures, for which the $d_{ii'}^2$ can serve as the required input, help to spot clusters of (one or more) points, a few of which may be composed of outliers. In an agglomerative-type procedure, such as those in Johnson [1967], the expectation is that outlier clusters will join the main body of points near or at the final level of clustering.

Example 5. In connection with the statistical study of groupings of corporations mentioned in Example 1, a principal components analysis was carried out on data which consisted of 14-dimensional observations on 32 chemical companies for the year 1965. The analysis was performed on the correlation matrix because the 14 variables involved were on widely differing scales. The striking feature in the plot of the first two principal components shown in Figure 5 (for which $c_1 = 5.81$ and $c_2 = 2.62$) is the presence of one clear-cut outlier, company 14.

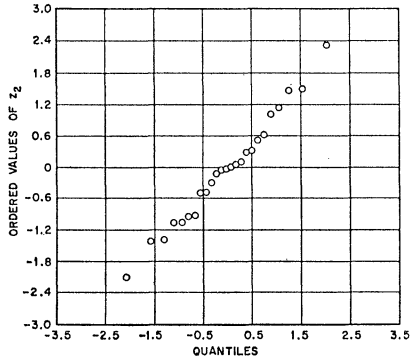

FIGURE 6b

NORMAL PROBABILITY PLOT OF SECOND PRINCIPAL COMPONENT IN EX. 6

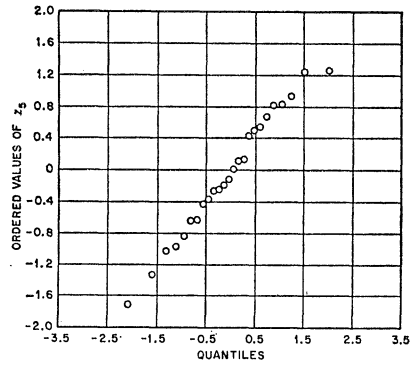

FIGURE 6c

NORMAL PROBABILITY PLOT OF FIFTH PRINCIPAL COMPONENT IN EX. 6

Example 6. The data in this case consist of 25 computer-generated observations from a 5-dimensional standard spherical normal distribution. Figures 6a, b, and c show typical normal probability plots for 3 of the 5 possible principal components of **S**. The configurations are reasonably linear with slopes quite close to the anticipated values. Next, an outlier was simulated by adding a large deviation to one of the 25 original observations. The normal probability plot of the projections onto the first principal component for the data with the outlier is exhibited in Figure 6d. The deviant observation (viz. the smallest one) is the projection of the outlier.

Example 7. To illustrate the use of generalized squared distances, the d_{i0}^2 and the $d_{ii'}^2$ were calculated, first of all, from a computer-generated sample of size 25 based on a standard spherical normal distribution and, a second time, from the same sample but with $(3, 3, 3)'$ added to $\mathbf{y}_1$ to simulate an outlier. Both gamma- and F-type probability plots were made for the 4 separate sets of distances, and it was found that the two distributions gave

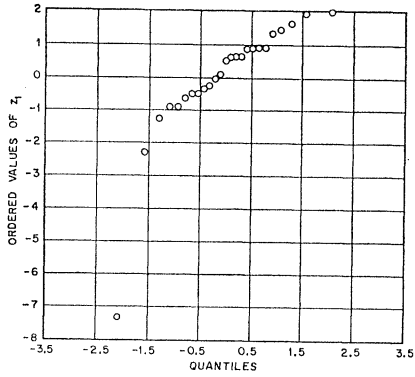

FIGURE 6d

NORMAL PROBABILITY PLOT OF FIRST PRINCIPAL COMPONENT WITH SIMULATED OUTLIER IN EX. 6

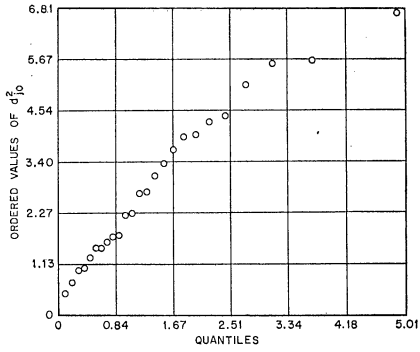

FIGURE 7a

GAMMA PROBABILITY PLOT OF 25 GENERALIZED SQUARED DISTANCES IN EX. 7 (SHAPE PARAMETER = 1.5)

virtually the same indications. Figures 7a and 7b show the gamma plots for the d^2_{i0}, and Figures 7c and 7d give the F plots for the transformed values of the $d^2_{ij'}$. The wild value at the top of Figure 7b corresponds to d^2_{10}. The 14 largest values in Figure 7d are all distances from $\mathbf{y}_1$. The remaining 10 of the d^2_{1j} are among the next 12 largest values in the figure. Two impressive features of Figures 7c and 7d are the zero intercept and the straight line configuration of the smaller $d^2_{ij'}$. The tendency of the points in Figures 7a and 7c to bend over at the top may be due in part to the correlations among the distances.

Example 8. The generalized squared interpoint distances between the chemical companies in Example 5 were used as input for a hierarchical cluster analysis. At each of the successive clustering levels, the two clusters with the smallest average (called the cluster value) of squared interpoint distances of companies in one existing cluster from those in another were joined together. The results along with the cluster values are displayed in Figure 8. The clustering begins with the joining of companies 11 and 31

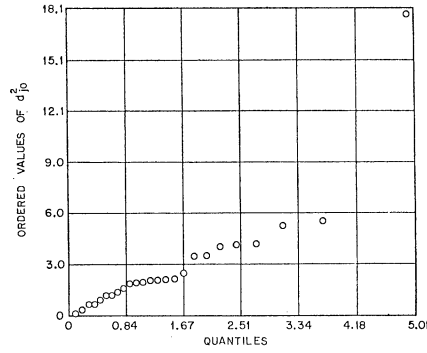

FIGURE 7b

GAMMA PROBABILITY PLOT OF 25 GENERALIZED SQUARED DISTANCES WITH SIMULATED OUTLIER IN EX. 7 (SHAPE PARAMETER = 1.5)

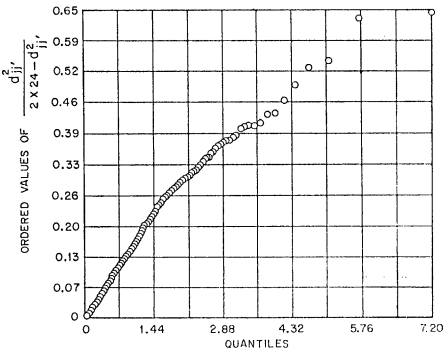

FIGURE 7c

F(3,21) PROBABILITY PLOT OF VALUES OBTAINED FROM 300 GENERALIZED SQUARED INTERPOINT DISTANCES IN EX. 7

and ends with company 14 merging with all of the others. Again company 14 appears to be an outlier.

4.2. *External analysis techniques*

The discriminant analysis of two or more groups of observations and the canonical analysis of two or more sets of variables are among the basic multivariate external analysis techniques. According to one method of discriminant analysis (cf. Rao [1965]), the key quantities are the positive eigenvalues, $f_1 \geq f_2 \geq \cdots \geq f_t$, and corresponding eigenvectors, $\{\mathbf{g}_1, \cdots, \mathbf{g}_t\}$, of $\mathbf{W}^{-1}\mathbf{B}$, where $t = \min(p, g - 1)$, p is the number of variables, g is the number of groups, $\mathbf{B}$ is the sample covariance matrix among group means, and $\mathbf{W}$ is the pooled within-groups sample covariance matrix. The eigenvectors are used to define projections $\mathbf{g}_i'\mathbf{Y}$, $i = 1, \cdots, t$, onto a t-dimensional discriminant space. The associated coordinates account, in progressively decreasing amounts measured by the f_i, for all of the differences among the group means.

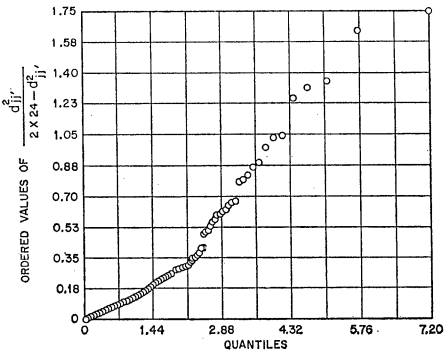

FIGURE 7d

F(3,21) PROBABILITY PLOT OF VALUES OBTAINED FROM 300 GENERALIZED SQUARED INTERPOINT DISTANCES WITH SIMULATED OUTLIER IN EX. 7

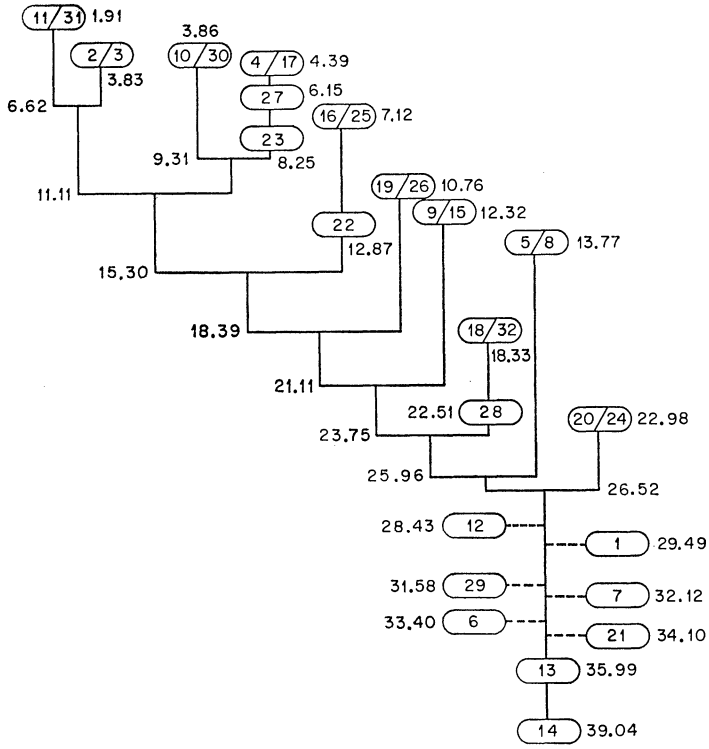

FIGURE 8

HIERARCHICAL CLUSTER ANALYSIS OF 32 CHEMICAL COMPANIES

A canonical analysis involves the formation of one or more stages of canonical variables which are selected to capture linear relationships among the sets of variables. The canonical variables are unit-variance linear compounds of the different sets, and at each stage one canonical variable is chosen from each set. In the classical two-set case (cf. Hotelling [1936]), the first stage or pair of canonical variables is the one with maximum or first canonical correlation $r^{(1)}$. The subsequent canonical pairs have maximum (canonical) correlations $r^{(2)}$, $r^{(3)}$, $\cdots$ subject to being uncorrelated with all of the preceding pairs. Five ways of defining canonical variables for more than two sets are given in Kettenring [1971].

Valuable insight can be gleaned from two- and three-dimensional displays of the discriminant and canonical variables. Such views of the discriminant space exhibit relative size, shape, and location of the groups as well as possible peculiarities in the positions of individual points. The discriminant analysis may be preceded by internal analyses of the individual groups for outliers with the hope of making the dispersions within the individual groups similar, as is required for the validity of the standard multigroup discriminant analysis procedure. The remaining observations can then be used to derive the discriminant coordinates, but the visual displays may profitably include

the positions of all of the data in the transformed space. The canonical variable plot, another mechanism for data exposure, can unearth outliers which are inducing an artificial linear relationship among the sets. Plots of the principal components (or other appropriate linear functions) of the canonical variables, as discussed in Kettenring [1971], are alternative summaries which have special appeal when the number of sets is large.

Normal probability plots and univariate outlier procedures can be applied to the canonical variables, or to linear functions of them, and to the discriminant variables, making a separate plot for each group. The slopes of the configurations on the last-mentioned of these plots provide a partial check on the homogeneity of dispersion among the groups.

Two examples of univariate statistics which are sensitive to the type of multivariate effects of interest in discriminant and canonical analyses are given below.

$$w_{kj}^2 = \sum_i f_i[\mathbf{g}_i'(\mathbf{y}_{kj} - \bar{\mathbf{y}}_k)]^2 = (\mathbf{y}_{kj} - \bar{\mathbf{y}}_k)'\mathbf{W}^{-1}\mathbf{B}\mathbf{W}^{-1}(\mathbf{y}_{kj} - \bar{\mathbf{y}}_k),$$

$$k = 1, \cdots, g, \quad j = 1, \cdots, n_k, \tag{23}$$

where $\mathbf{y}_{kj}$ is the jth observation in the kth group, $\bar{\mathbf{y}}_k$ is the kth group mean and n_k is the number of observations in the kth group. w_{kj}^2 is a weighted sum of squares of the projections of $\mathbf{y}_{kj} - \bar{\mathbf{y}}_k$ onto the discriminant axes, and $\sum\sum w_{kj}^2 = (n - g)\sum f_i$.

$$x_j^2 = \prod_i (1 - r^{(i)^2})/\prod_i (1 - r_{-j}^{(i)^2})$$

$$= \frac{1 - \left(\frac{n}{n-1}\right)(\mathbf{y}_j - \bar{\mathbf{y}})'\mathbf{A}^{-1}(\mathbf{y}_j - \bar{\mathbf{y}})}{\left[1 - \left(\frac{n}{n-1}\right)(\mathbf{y}_{1j} - \bar{\mathbf{y}}_1)'\mathbf{A}_{11}^{-1}(\mathbf{y}_{1j} - \bar{\mathbf{y}}_1)\right]\left[1 - \left(\frac{n}{n-1}\right)(\mathbf{y}_{2j} - \bar{\mathbf{y}}_2)'\mathbf{A}_{22}^{-1}(\mathbf{y}_{2j} - \bar{\mathbf{y}}_2)\right]},$$

$$j = 1, \cdots, n, \tag{24}$$

where

$$\mathbf{y}_j' = (\mathbf{y}_{1j}' \mid \mathbf{y}_{2j}'), \qquad \bar{\mathbf{y}}' = (\bar{\mathbf{y}}_1' \mid \bar{\mathbf{y}}_2') \quad \text{and} \quad \mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

are partitioned in accordance with the dimensions of the two sets. (The subscript which designated the group in (23) now refers to the set.)

Some possible aids for examining these statistics, which need further investigation, include gamma probability plots of the w_{kj}^2 both for a specific group (i.e. fixed k) and across all groups, and a normal probability plot of the values of

$$\log x_j^2 = \sum_i \left[\log\left(\frac{1 + r^{(i)}}{1 - r_{-j}^{(i)}}\right) - \log\left(\frac{1 + r_{-j}^{(i)}}{1 - r^{(i)}}\right)\right]. \tag{25}$$

Example 9. As part of the study mentioned in Example 1, four groups of companies—chemicals, drugs, oils, and steels—were subjected to a dis-

criminant analysis using 14-dimensional observations on each of 92 companies for the year 1965. Figure 9a shows the points projected onto the space of the first two of the three discriminant variables. The eigenvalues associated with the three variables are $f_1 = 87.32$, $f_2 = 16.53$, and $f_3 = 7.96$. Internal examination of the groups uncovered several possible outliers and in a follow-up analysis these companies were not used in the calculation of the discriminant coordinates and eigenvalues. Figure 9b is a display of all of the points in the revised two-dimensional discriminant space. A few of the outlier points fall near their original groups, a few fall near other groups, and 4 fall outside the dimensions of the figure. In the last category, the one from the A group corresponds to the outlier detected in Examples 5 and 6. The group separation is now somewhat greater: $f_1 = 109.36$, $f_2 = 10.94$, and $f_3 = 7.54$. Even with outliers excluded, there is still empirical evidence in the figure of nonhomogeneity of dispersion.

5. CONCLUDING REMARKS

This paper has attempted to survey and to suggest some problems and procedures which are important in the analysis of multiresponse data. Methods for protecting statistics that summarize multivariate location and dispersion from possible outliers, as well as techniques for aiding the detection of such outliers, were included in the coverage.

Many issues have been raised and need further study before they can be resolved satisfactorily. Some of these have been discussed in the preceding sections but others may deserve to be raised anew or reiterated.

For instance, the question of the commutativity of robust estimators

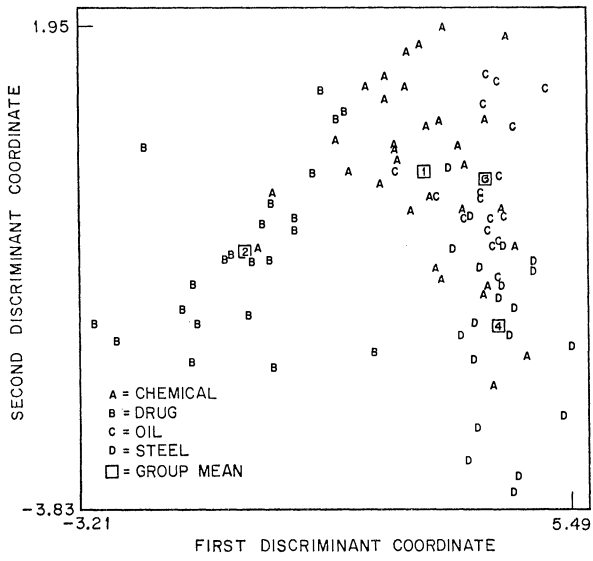

FIGURE 9a

FIRST TWO DISCRIMINANT COORDINATES FOR 4 GROUPS BASED ON ALL COMPANIES

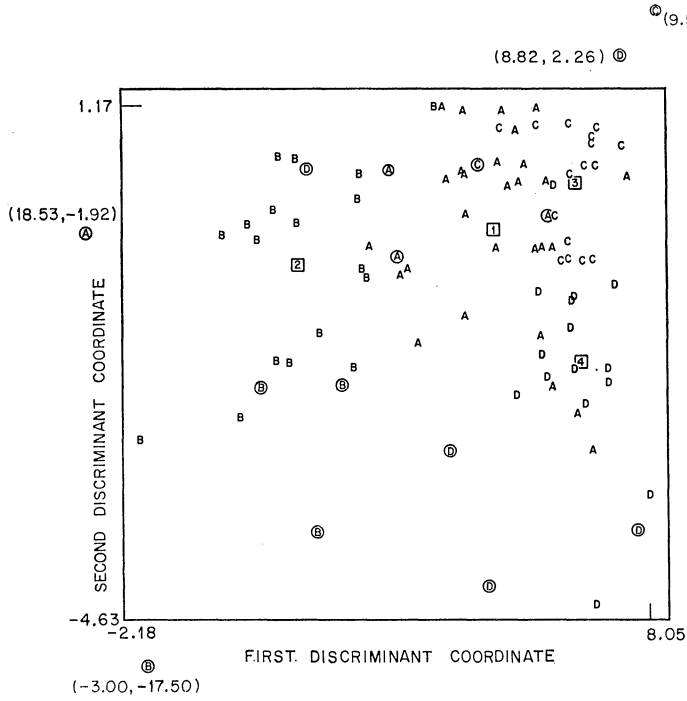

FIGURE 9b

REVISED DISCRIMINANT COORDINATES (CIRCLED LETTERS DESIGNATE COMPANIES TREATED AS OUTLIERS)

of multivariate location may be approached by applying the estimators considered in section 2.1 after a preliminary transformation to principal components coordinates or to sphericized (i.e. transformed by $S^{-\frac{1}{2}}$) coordinates. With a data-dependent transformation, viz. when the sample covariance matrix or correlation matrix is used for obtaining the transformation, one may wish to robustify the transformation. Thus, for instance, an initial robust estimate of the covariance matrix may be obtained from pairwise differences of observations as described in section 2.2.3, and this estimate may then be used for determining the principal components transformation or the sphericizing transformation.

A second issue deserving comment is that, while the very limited Monte Carlo results discussed in section 2 are all concerned with the null 'normal' case, it is clearly important that the performance of the proposed robust estimates under several alternatives needs to be investigated.

Another question is posed by the methods proposed in section 2.2.3 for developing positive definite robust estimators of the covariance matrix—is it possible to develop such estimators without relying on the number of observations having to exceed the dimensionality of response?

A forth issue arises from the intertwining of the objectives of techniques for detecting multivariate outliers with those related to methods of assessing

the joint normality of multiresponse data. Some of the techniques developed by Andrews *et al.* [1971b] for the latter purpose might also be applicable for detecting multivariate maverick observations.

It should be reemphasized that many of the methods suggested for the first time in this paper need more detailed study and development. From details, such as the multiplicative constants required to make the robust estimators of the covariance matrices unbiased, to ones that pertain to the appropriate distribution theory of the statistics suggested in section 4 for delineating outliers, more work is needed. The authors are currently involved in pursuing some of this work.

ESTIMEES ROBUSTES, RESIDUS ET DETECTION D'OBSERVATIONS ABERRANTES AVEC DES DONNEES A PLUSIEURS REPONSES

RESUME

L'article donne un aperçu des concepts et des techniques se rattachant à (i) l'estimation robuste de la position et de la dispersion multivariates; (ii) l'analyse de deux types de résidus multidimensionnels—à savoir ceux qui interviennent dans le contexte de l'analyse en composantes principales aussi bien que les résidus plus connus associés à l'ajustement par les moindres carrés; et (iii) la détection des observations aberrantes à plusieurs réponses. l'accent est porté sur les méthodes d'analyse exploratoire et on englobe à la fois une revue des techniques qui existent et un essai pour proposer, timidement, une méthodologie nouvelle qui demande encore investigation et développement. On donne quelques exemples d'utilisation de ces méthodes.

ACKNOWLEDGMENT

The authors are thankful to Mrs. Elisa Lee for her assistance in obtaining the results of the Monte Carlo study of robust estimators summarized in section 2.

REFERENCES

Abrahamson, I. G., Gentleman, J. F., Gnanadesikan, R., Walcheski, A. F., and Williams, D. E. [1969]. Statistical methods for studying aging and for selecting semiconductor devices. *ASQC Tech. Conf. Trans.*, 533–40.

Andrews, D. F. [1971]. Significance tests based on residuals. *Biometrika 58*, 139–48.

Andrews, D. F., Bickel, P. J., Hampel, F., Huber, P. J., Rogers, W. H., and Tukey, J. W. [1971a]. Robust estimates of location: survey and advances. Unpublished manuscript.

Andrews, D. F., Gnanadesikan, R., and Warner, J. L. [1971b]. Methods for assessing multivariate normality. Unpublished manuscript.

Anscombe, F. J. [1960]. Rejection of outliers. *Technometrics 2*, 123–47.

Anscombe, F. J. [1961]. Examination of residuals. *Proc. 4th Berkeley Symp. Math. Statist. Prob. I*, 1–36.

Anscombe, F. J. and Tukey, J. W. [1963]. The examination and analysis of residuals. *Technometrics 5*, 141–60.

Bickel, P. J. [1964]. On some alternative estimates for shift in the p-variate one sample problem. *Ann. Math. Statist. 35*, 1079–90.

Chen, H. J., Gnanadesikan, R., Kettenring, J. R., and McElroy, M. B. [1970]. A statistical study of groupings of corporations. *1970 Business and Econ. Statist. Sec. Proc., Amer. Statist. Ass.*, 447–51.

Cox, D. R. [1968]. Notes on some aspects of regression analysis. *J. R. Statist. Soc. A 131*, 265–79.

Cox, D. R. and Snell, E. J. [1968]. A general definition of residuals. *J. R. Statist. Soc. B 30*, 248–75.

Daniel, C. [1960]. Locating outliers in factorial experiments. *Technometrics 2*, 149–56.

David, H. A. [1970]. *Order Statistics*. Wiley, New York.

Dixon, W. J. [1960]. Simplified estimation from censored normal samples. *Ann. Math. Statist. 31*, 385–91.

Gentleman, W. M. [1965]. Robust estimation of multivariate location by minimizing pth power deviations. Unpublished Ph.D. thesis, Princeton University.

Gnanadesikan, R., Pinkham, R. S., and Hughes, L. P. [1967]. Maximum likelihood estimation of the parameters of the beta distribution from smallest order statistics. *Technometrics 9*, 607–20.

Gnanadesikan, R. and Wilk, M. B. [1968]. Methods for reduction of dimensionality in multiresponse data. Unpublished manuscript.

Gnanadesikan, R. and Wilk, M. B. [1969]. Data analytic methods in multivariate statistical analysis. In: *Multivariate Analysis II*. Krishnaiah, P. R. (Ed.), Academic Press, New York. 593–638.

Hampel, F. R. [1968]. Contributions to the theory of robust estimation. Unpublished Ph.D. thesis, Univ. of Calif., Berkeley.

Healy, M. J. R. [1968]. Multivariate normal plotting. *Appl. Statist. 17*, 157–61.

Hodges, J. L. and Lehmann, E. L. [1963]. Estimates of location based on rank tests. *Ann. Math. Statist. 34*, 598–611.

Hotelling, H. [1936]. Relations between two sets of variates. *Biometrika 28*, 321–77.

Huber, P. J. [1964]. Robust estimation of a location parameter. *Ann. Math. Statist. 35*, 73–101.

Huber, P. J. [1970]. Studentizing robust estimates. In: *Nonparametric Techniques in Statistical Inference*. Puri, M. L. (Ed.), Cambridge Univ. Press. 453–63.

Jaeckel, L. A. [1971]. Robust estimates of location: symmetry and asymmetric contamination. *Ann. Math. Statist. 42*, 1020–34.

Johnson, N. L. and Leone, F. C. [1964]. *Statistics and Experimental Design in Engineering and the Physical Sciences, Vol. I*. Wiley, New York.

Johnson, S. C. [1967]. Hierarchical clustering schemes. *Psychometrika 32*, 241–54.

Kettenring, J. R. [1971]. Canonical analysis of several sets of variables. To appear in *Biometrika 58*.

Larsen, W. A. and McCleary, S. J. [1970]. The use of partial residual plots in regression analysis. Unpublished manuscript. (Submitted to *Technometrics*.)

McLaughlin, D. H. and Tukey, J. W. [1961]. The variance of means of symmetrically trimmed samples from normal populations, and its estimation from such trimmed samples. (Trimming/Winsorization I.) Tech. Report 42, Statist. Tech. Res. Group, Princeton Univ.

Mood, A. M. [1941]. On the joint distribution of the medians in samples from a multivariate population. *Ann. Math. Statist. 12*, 268–78.

Pearson, K. [1901]. On lines and planes of closest fit to systems of points in space. *Phil. Mag. 2*, (Series VI), 559–72.

Rao, C. R. [1964]. The use and interpretation of principal component analysis in applied research. *Sankhyā A 26*, 329–58.

Rao, C. R. [1965]. *Linear Statistical Inference and its Applications*. Wiley, New York.

Roy, S. N., Gnanadesikan, R., and Srivastava, J. N. [1971]. *Analysis & Design of Certain Quantitative Multiresponse Experiments*. Pergamon Press, Oxford.

Siotani, M. [1959]. The extreme value of the generalized distances of the individual points in the multivariate normal sample. *Ann. Inst. Statist. Math., Tokyo, 10*, 183–203.

Teichroew, D. [1956]. Tables of expected values of order statistics and products of order statistics for samples of size twenty and less from the normal distribution. *Ann. Math. Statist. 27*, 410–26.

Terry, M. E. [1955]. On the analysis of planned experiments. *National Convention ASQC Trans.*, 553–6.

Tukey, J. W. [1960]. A survey of sampling from contaminated distributions. In: *Contributions to Probability and Statistics* Olkin, I. (Ed.), Stanford Univ. Press. 448–85.

Tukey, J. W. [1962]. The future of data analysis. *Ann. Math. Statist. 33*, 1–67.

Tukey, J. W. [1970]. *Exploratory Data Analysis.* Limited preliminary edition, Addison-Wesley, Reading.

Tukey, J. W. [1971]. Personal communication.

Tukey, J. W. and McLaughlin, D. H. [1963]. Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization 1. *Sankhyā A 25*, 331–52.

Tukey, J. W. and Wilk, M. B. [1966]. Data analysis and statistics: An expository overview. *AFIPS Conf. Proc., Fall Joint Comp. Conf. 29*, 695–709.

Wilk, M. B. and Gnanadesikan, R. [1964]. Graphical methods for internal comparisons in multiresponse experiments. *Ann. Math. Statist. 35*, 613–31.

Wilk, M. B., Gnanadesikan, R., and Huyett, M. J. [1962a]. Probability plots for the gamma distribution. *Technometrics 4*, 1–20.

Wilk, M. B., Gnanadesikan, R., and Huyett, M. J. [1962b]. Estimation of parameters of the gamma distribution using order statistics. *Biometrika 49*, 525–45.

Wilk, M. B., Gnanadesikan, R., Huyett, M. J., and Lauh, E. [1962]. A study of alternate compounding matrices used in a graphical internal comparisons procedure. Unpublished manuscript.

Wilks, S. S. [1962]. *Mathematical Statistics.* Wiley, New York.

Wilks, S. S. [1963]. Multivariate statistical outliers. *Sankhyā A 25*, 407–26.

Received August 1971, *Revised September* 1971

Key Words: Robust estimation of multivariate location and dispersion; Analysis of usual and robustified least squares residuals and principal components residuals; Various statistics for detecting multivariate outliers.