Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation
Author(s): N. A. Campbell
Source: *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol. 29, No. 3 (1980), pp. 231-237
Published by: Wiley for the Royal Statistical Society
Stable URL: https://www.jstor.org/stable/2346896
Accessed: 15-10-2018 06:31 UTC

*Appl. Statist.* (1980),
**29**, *No.* 3, *pp.* 231–237

# Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation

By N. A. CAMPBELL†

*Imperial College, London, UK*

[Received November 1978. Final revision March 1980]

SUMMARY

The detection of atypical observations from multivariate data sets can be enhanced by examining probability plots of Mahalanobis squared distances using robust $M$-estimates of means and of covariances, rather than the usual maximum likelihood estimates. The weights associated with the robust estimation can also be used to indicate atypical observations. For uncontaminated data, the robust estimates are similar to the usual estimates.

A procedure for robust principal component analysis is given; it also indicates atypical observations and provides an analysis relatively little influenced by such observations.

*Keywords*: ROBUST ESTIMATION; M-ESTIMATORS; OUTLIER DETECTION; PRINCIPAL COMPONENT ANALYSIS; MULTIVARIATE NORMALITY

## 1. INTRODUCTION

OBSERVATIONS which are grossly atypical in a single component can often be detected by applying univariate techniques to each variable. For multivariate data, observations are often only found to be atypical when the value for each variable is considered in relation to the other variables; some values fail to maintain the pattern of relationships between the variables evident in the majority of the observations. Since the performance of classical procedures is seriously influenced by atypical values, robust methods which are little influenced by atypical values provide an attractive complementary approach. The survey papers by Huber (1972) and Hampel (1973) summarize the important methods and results from the earlier years of univariate robust studies, while Hampel (1977) and Huber (1977b) include a review of more recent results. An introductory paper by Hogg (1977) gives the basic ideas.

The emphasis throughout this paper is on the provision of estimates of means and of covariances for a single group which are little influenced by atypical observations, and on the detection of observations having undue influence on the estimates. The underlying distribution is assumed, after transformation if necessary, to be symmetric; a multivariate Gaussian form is examined in the probability plotting. Gnanadesikan (1977, Chapter 5) discusses transformations to achieve approximate symmetry.

Robust $M$-estimation of means and covariances is reviewed in Section 2, and its use in conjunction with probability plots of associated Mahalanobis squared distances is considered. A procedure for robust principal component analysis is proposed in Section 3. Typical data sets are examined in Section 4, while some general recommendations are given in Section 5.

## 2. ROBUST ESTIMATION OF MULTIVARIATE LOCATION AND SCATTER

Healy (1968) and Cox (1968) have suggested an extension of probability plots of univariate data to the multivariate situation, by plotting the Mahalanobis squared distance of each observation against the order statistic for a chi-squared distribution with $v$ d.f., where $v$ is the

† Now at Division of Mathematics and Statistics, CSIRO, Western Australia.

number of variables (see also Gnanadesikan, 1977, p. 172). If $\bar{\mathbf{x}}$ represents the $v \times 1$ vector of sample means, and $\mathbf{V}$ the sample covariance matrix, then the Mahalanobis squared distance of the $m$th observation from the mean of the observations is defined by $d_m^2 = (\mathbf{x}_m - \bar{\mathbf{x}})^T \mathbf{V}^{-1} (\mathbf{x}_m - \bar{\mathbf{x}})$, $m = 1, \ldots, n$. The approach combines examination of the distributional assumption of a multivariate Gaussian form with detection of atypical observations and is especially suited to informal graphical description.

Atypical multivariate vectors of observations will tend to deflate correlations and possibly inflate variances. This will in general decrease the Mahalanobis distance for the atypical observation and distort the rest of the plot.

The Mahalanobis distances play a basic role in multivariate $M$-estimation (see Maronna, 1976; Hampel, 1977; Huber, 1977a, b). The estimators can be derived by assuming an elliptically symmetric density, and then associating this with a contaminated Gaussian density.

From the applied viewpoint, $M$-estimators can be considered as a simple modification of classical estimators; they give full weight to observations assumed to come from the main body of the data, but reduced weight or influence to observations from the tails of the contaminating distribution. In practice, the influence of observations with unduly large Mahalanobis distances is downweighted.

The equations used here to define robust estimators of means and covariances are as follows:

$$\bar{\mathbf{x}} = \sum_{m=1}^{n} w_m \mathbf{x}_m \Big/ \sum_{m=1}^{n} w_m \tag{1}$$

and

$$\mathbf{V} = \sum_{m=1}^{n} w_m^2 (\mathbf{x}_m - \bar{\mathbf{x}})(\mathbf{x}_m - \bar{\mathbf{x}})^T \Big/ \left( \sum_{m=1}^{n} w_m^2 - 1 \right), \tag{2}$$

where

$$w_m = w(d_m) = \omega(d_m)/d_m$$

and

$$d_m = \{(\mathbf{x}_m - \bar{\mathbf{x}})^T \mathbf{V}^{-1} (\mathbf{x}_m - \bar{\mathbf{x}})\}^{\frac{1}{2}}.$$

The solution for $\bar{\mathbf{x}}$ and $\mathbf{V}$ is iterative.

Here $\omega$ is a bounded influence function (Hampel, 1974), often linear over the range of values of $d_m$ corresponding to reasonable data, but bounded outside this range. Hampel (1973) has suggested that the influence and hence the weight of an extreme atypical observation should be zero, so that $\omega$ should redescend for sufficiently large values of $d_m$.

The two-parameter form of $\omega$ used here is

$$\begin{aligned} \omega(d) &= d \quad \text{if } d \leqslant d_0 \\ &= d_0 \exp\{-\tfrac{1}{2}(d - d_0)^2 / b_2^2\} \quad \text{if } d > d_0, \end{aligned} \tag{3}$$

where

$$d_0 = \sqrt{v} + b_1/\sqrt{2}.$$

The motivation behind this form of $d_0$ is that, under standard assumptions, $d^2 \sim \chi_v^2$, and so Fisher's square root approximation gives $d \sim N(\sqrt{v}, 1/\sqrt{2})$; $b_1$ is equated with a percentage point of the standard Gaussian distribution. The parameter $b_2$ controls the rate of decrease of the influence function to zero.

Following extensive empirical experience, it seems sufficient for practical applications to consider three members of (3). They are

(a) $b_1 = \infty$ — all observations are given unit weight, so the usual estimates result;
(b) $b_1 = 2, b_2 = \infty$ — the non-descending form suggested by Huber (1964);
(c) $b_1 = 2, b_2 = 1{\cdot}25$ — a redescending form, giving the qualitative behaviour suggested by Hampel (1973), so that the weight function decreases at a faster rate than in (b).

These recommended values of $b_1$ and $b_2$ are further supported by asymptotic results and results from the analysis of some generated data.

If the robust estimates are to be used in subsequent statistical analysis, it is important that they differ little from the usual estimates when applied to uncontaminated data. Table 1 gives a stem-and-leaf plot of the ratio of robust to usual estimates of variance for ten sets of generated multivariate Gaussian data with 7 variables; each set consists of six groups with sample size and underlying means and covariances corresponding to the first six groups of the *Thais* data considered in Section 4. It seems reasonable to conclude from Table 1 that the robust estimates

TABLE 1

*Stem-and-leaf plot for ratios of robust to usual variances for generated multivariate Gaussian data with same underlying structure as* Thais *data (7 variables, 60 groups).*

| | |
|---|---|
| 1·01 | 1[a], 2, 2, 3, 3 |
| 1·00 | 0(243)[b], 1(19), 2(9), 3, 3, 4, 4, 5, 6, 6, 6, 7 |
| 0·99 | 0(9), 1(6), 2, 3(10), 4(7), 5(5), 6(7), 7(7), 8(13), 9(19) |
| 0·98 | 1, 1, 2, 2, 4, 5, 5, 6, 6, 6, 8, 8, 8, 9, 9, 9 |
| 0·97 | 3, 4, 4, 7, 8, 8 |
| 0·96 | 1, 2, 2, 2, 7, 9 |
| | 0·952, 0·954, 0·957, 0·947, 0·934, 0·934, 0·938, 0·927, 0·928, 0·929, 0·910, 0·912, 0·912, 0·89, 0·88, 0·88, 0·87, 0·87, 0·86, 0·82 |

[a] Value is 1·011

[b] Value 1·000 is repeated 243 times.

of the variances will generally be within 2–3 per cent of the usual estimates for well-behaved data. For multivariate Gaussian data, multiplicative correction factors can be calculated to give large sample agreement of the expected value of $d^2$ under robust and usual estimation. Asymptotically, $E(d^2) = v$ if $b_1 = \infty$. For the non-descending influence function,

$$E(d^2) = v\,\mathrm{Ch}(d_0^2;\, v+2) + d_0^2\{1-\mathrm{Ch}(d_0^2;\, v)\},$$

where $\mathrm{Ch}(.;\, v)$ denotes the cumulative chi-squared distribution function on $v$ d.f. For the redescending influence function,

$$E(d^2) = v\,\mathrm{Ch}(d_0^2;\, v+2) + d_0^2 \int_{d_0^2}^{\infty} \exp\{-\tfrac{1}{2}(d-d_0)^2/b_2^2\}\, f_v(d^2)\, d(d^2)$$

where $f_v(d^2)$ is the chi-squared density on $v$ d.f. With $b_1 = 2{\cdot}0$ and $b_2 = 1{\cdot}25$, the asymptotic correction factor is less than 1·020 for the non-descending influence function and less than 1·025 for the redescending function.

## 3. Robust Principal Component Analysis

A principal component analysis of the usual sample covariance matrix $\mathbf{V}$ (or associated correlation matrix $\mathbf{R}$) involves finding the linear combination $y_m = \mathbf{u}^{\mathrm{T}}\mathbf{x}_m$ of the original variables $\mathbf{x}_m$ such that the usual sample variance of the $y_m$ is a maximum. The solution is given by an eigenanalysis of $\mathbf{V}$, viz. $\mathbf{V} = \mathbf{UEU}^{\mathrm{T}}$. The eigenvectors $\mathbf{u}_i$, given by the columns of $\mathbf{U}$, define the linear combinations, while the corresponding diagonal elements $e_i$ of the diagonal matrix of eigenvalues $\mathbf{E}$ are the sample variances of the derived variables.

An obvious way of modifying the analysis is to replace $\mathbf{V}$ by the robust estimator in (2); this is the $M$-estimator solution to robust principal components. An observation is weighted according to its total distance $d_m$ from the robust estimate of location. This distance can be decomposed into components along each eigenvector; and an observation may have a large component along one direction and small components along the remaining directions and

hence not be adequately downweighted. Robust $M$-estimation of mean and variance can be applied to each principal component, to determine directions which are little influenced by atypical observations.

The proposed procedure is as follows:

1. Take as an initial estimate of $\mathbf{u}_1$ the first eigenvector from an eigenanalysis of $\mathbf{V}$.
2. Form the principal component scores $y_m = \mathbf{u}_1^{\mathrm{T}} \mathbf{x}_m$.
3. Determine the $M$-estimators of mean and variance of $y_m$, and the associated weights $w_m$. The median and $\{0{\cdot}74\,(\text{interquartile range})\}^2$ of the $y_m$ can be used to provide initial robust estimates. Here $0{\cdot}74 = (2 \times 0{\cdot}675)^{-1}$ and $0{\cdot}675$ is the 75 per cent quantile for the $N(0,1)$ distribution. This initial choice ensures that the proportion of observations downweighted is kept reasonably small.

3(a). After the first iteration, take the weights $w_m$ as the minimum of the weights for the current and previous iterations; this prevents oscillation of the solution.

4. Calculate $\bar{\mathbf{x}}$ and $\mathbf{V}$ as in (1) and (2) using the weights $w_m$ from stage 3.
5. Determine the first eigenvalue and eigenvector $\mathbf{u}_1$ of $\mathbf{V}$.
6. Repeat steps (2) to (5) until successive estimates of the eigenvalue are sufficiently close.

To determine successive directions $\mathbf{u}_i$, $2 \leqslant i \leqslant v$, project the data onto the space orthogonal to that spanned by the previous eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_{i-1}$, and repeat steps (2) to (5); take as the initial estimate the second eigenvector from the last iteration for the previous eigenvector. The proposed procedure for successive directions can be set out as follows.

7. Form $\mathbf{x}_{im} = (\mathbf{I} - \mathbf{U}_{i-1}\,\mathbf{U}_{i-1}^{\mathrm{T}})\,\mathbf{x}_m$, where $\mathbf{U}_{i-1} = (\mathbf{u}_1, \ldots, \mathbf{u}_{i-1})$.
8. Repeat steps (2) to (5) with $\mathbf{x}_{im}$ replacing $\mathbf{x}_m$, and determine the first eigenvector $\mathbf{u}$.
9. The principal component scores are given by $\mathbf{u}^{\mathrm{T}}\,\mathbf{x}_{im} = \mathbf{u}^{\mathrm{T}}(\mathbf{I} - \mathbf{U}_{i-1}\,\mathbf{U}_{i-1}^{\mathrm{T}})\,\mathbf{x}_m$ and hence $\mathbf{u}_i = (\mathbf{I} - \mathbf{U}_{i-1}\,\mathbf{U}_{i-1}^{\mathrm{T}})\,\mathbf{u}$.

Steps (7), (8) and (9) are repeated until all $v$ eigenvalues $e_i$ and eigenvectors $\mathbf{u}_i$, together with the associated weights, are determined. Alternatively the procedure may be terminated after some specified proportion of variation is explained.

Finally, a robust estimate of the covariance or correlation matrix can be found from $\mathbf{UEU}^{\mathrm{T}}$, to provide an alternative robust estimate. Both this approach and that described in the previous Section give a positive definite correlation/covariance matrix. Robust estimation of each entry separately does not always achieve this.

## 4. Some Practical Examples

The first data set to be discussed is taken from a study of geographic variation in the whelk *Thais lamellosa* (Campbell, 1978). Data are available for twelve groups on the west coast of North America. The numbers in each group are: 50, 72, 99, 76, 37, 36, 46, 46, 51, 34, 28 and 43. Measurements were made on twenty variables; canonical variate analyses show that seven of these provide much of the between-groups discrimination. Robust covariance estimation was applied to the twelve groups based on the seven variables, and Gamma $(\frac{1}{2}; v/2)$ quantile–quantile $(Q$–$Q)$ probability plots of the associated $d_m^2$ and Gaussian $Q$–$Q$ probability plots of the $d_m^{2/3}$ were made. The latter uses the Wilson–Hilferty (1931) transformation of a gamma variate to Gaussian form (see also Healy, 1968, p. 159).

Probability plots of the distances derived with $b_1 = 2, b_2 = 1{\cdot}25$ show that nine of the groups have one or two atypical observations, while one group has several atypical observations. The associated weights are all less than $0{\cdot}35$; 18 of the 21 atypical observations have zero weight (to two decimal places). Probability plots of the usual distances indicate only ten of the atypical observations, with a further four or five doubtful.

Four of the 21 atypical observations have two atypical values in the vector, giving 25 out of a total of 4326 ($= 618 \times 7$ variables) variable values which are atypical. The variables for each group show high correlations; if the observations are listed in increasing order of the largest variable (here overall length), atypical values are readily apparent once the observation is

indicated. Corrections of either 100 or 50 units would provide good agreement with the values for adjacent observations for all but four values; interchanging the order of two numbers would explain a further two. The correction of 100 or 50 units is an acceptable one since the dial on the calipers used to measure the whelks records to 50 units. In more than 4000 measurements the linear scale will occasionally be misread.

Fig. 1(a) shows a Gaussian probability plot of usual distances for group 3 ($n = 99$) (i.e. a $Q$–$Q$ probability plot of cube root of squared Mahalanobis distances against Gaussian order

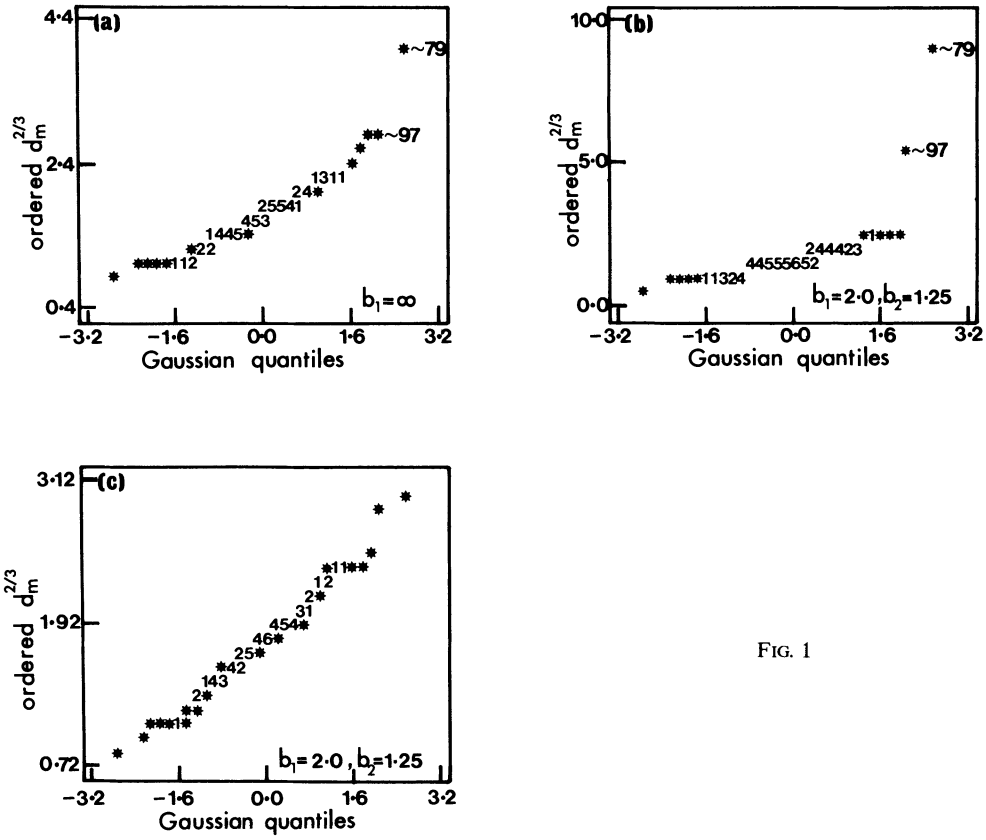


FIG. 1

statistics with the distances calculated using the usual means and covariances); one atypical observation (79) is indicated. Robust $M$-estimation with $b_1 = 2{\cdot}0, b_2 = \infty$ gives a weight of 0·03 for this observation; a second observation (97) has a weight of 0·11. With $b_2 = 1{\cdot}25$, both observations have a zero weight. Fig. 1(b) shows a Gaussian probability plot of robust distances when $b_1 = 2{\cdot}0, b_2 = 1{\cdot}25$. Two observations, 79 and 97, are clearly atypical. Both are atypical in the second variable, the first being out by 100 units and the second by 50 units. A Gaussian probability plot of robust distances for the corrected data (Fig. 1(c)), with $b_1 = 2{\cdot}0, b_2 = 1{\cdot}25$, is now linear. None of the weights is now less than 0·35. The usual estimate of standard deviation is 48·9; the robust estimate is 44·9 while the robust estimate after correcting the data is 45·8. The correlations of the second variable with the remaining variables are increased by 0·02 to 0·06.

The second data set is taken from an unpublished study of morphometric divergence in male and female scorpions occurring in Australia. Nine variables were measured on each specimen. The male data for one species ($n = 181$) are discussed here. Robust covariance estimation ($b_1 = 2{\cdot}0, b_2 = 1{\cdot}25$) indicates five observations with weights less than 0·35. A robust principal

component analysis indicates a further three observations with weights less than 0·10 for at least one component and a further sixteen with weights less than 0·35. The advantage of the combined approach is that different combinations of variables are examined. For example, one of the observations with a zero weight for robust covariance estimation has zero weight for principal components 5 and 6. Examination of the eigenvectors shows variable 8 to have a high loading for these components. The measurement of 3·5 was checked by my co-worker and found to be 6·0 (the robust estimate of standard deviation is 0·77). Another observation (not yet rechecked) has a weight of 0·60 for robust covariance estimation ($b_1 = 2{\cdot}0$, $b_2 = 1{\cdot}25$) but weights of 0·06, 0·21, 0·27 and 0·23 for principal components 3–6. Of those observations checked and corrected, the common error was a measurement out by 1 or 1·5 mm; the metal dial calipers were graduated to 0·05 mm.

## 5. Discussion

A recommended practical approach is to determine the means, covariances, distances and associated weights for $b_1 = \infty$; for $b_1 = 2{\cdot}0$, $b_2 = \infty$; and for $b_1 = 2{\cdot}0$, $b_2 = 1{\cdot}25$. Gaussian probability plots of cube root of squared distances (with $b_1 = 2{\cdot}0$, $b_2 = 1{\cdot}25$) will indicate atypical observations. For more than six or seven variables, a robust principal component analysis is also useful for identifying atypical observations. The weights $w_m^2$ associated with the $d_m^2$ also indicate atypical observations. Examination of a number of data sets and of the generated Gaussian data in Section 2 shows that a weight of less than 0·30 with $b_1 = 2{\cdot}0$, $b_2 = 1{\cdot}25$ (corresponding approximately to a weight of less than 0·60 with $b_2 = \infty$) always indicates an atypical observation. A weight of more than 0·70 with $b_2 = \infty$ is associated with a reasonable observation.

The Mahalanobis squared distances are usually plotted against the quantiles of a gamma distribution with shape parameter $v/2$, though the quantiles of a beta $\{v/2, (n-v-1)/2\}$ distribution may be more appropriate (Small, 1978). There is good agreement between $Q$–$Q$ plots of $d_m^2$ versus Gamma $(\frac{1}{2}, v/2)$ and of $d_m^{2/3}$ versus $N(0, 1)$, though the cube root transformation tends to lessen the visual impact of the large distances on the gamma plot. The general conclusion here reinforces the remarks of Healy (1968, p. 159) that the Gaussian plot seems more than adequate for detecting atypical observations and examining the multivariate Gaussian assumption.

## Acknowledgements

I am grateful to F. R. Hampel for stimulating discussions of some of the problems in robust estimation, and to M. J. R. Healy, D. R. Cox and a referee for their suggestions and comments on an earlier draft of the paper.

The author was supported by a CSIRO Divisional Postgraduate Research Studentship.

## References

Campbell, C. A. (1978). The frilled dogwinkle: ecological genetics of a morphologically variable snail, *Thais lamellosa*. Unpublished Ph.D. thesis, University of California, Davis.

Cox, D. R. (1968). Notes on some aspects of regression analysis. *J. R. Statist. Soc.* A, **131**, 265–279.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. New York: Wiley.

Hampel, F. R. (1973). Robust estimation: a condensed partial survey. *Z. Wahr, verw. Geb.*, **27**, 87–104.

—— (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Ass.*, **69**, 383–393.

—— (1977). Modern trends in the theory of robustness. Research Report No. 13, Swiss Federal Institute of Technology, Zurich.

Healy, M. J. R. (1968). Multivariate normal plotting. *Appl. Statist.*, **17**, 157–161.

Hogg, R. V. (1977). An introduction to robust procedures. *Comm. Statist. Theor. Meth.*, **A6**, 789–794.

Huber, P. J. (1964). Robust estimation for a location parameter. *Ann. Math. Statist.*, **35**, 73–101.

—— (1972). Robust statistics: a review. *Ann. Math. Statist.*, **43**, 1041–1067.

—— (1977a). Robust covariances. In *Statistical Decision Theory and Related Topics II* (S. S. Gupta and D. S. Moore, eds), pp. 165–191. New York: Academic Press.

—— (1977b). *Robust Statistical Procedures*. Philadelphia: SIAM.

MARONNA, R. A. (1976). Robust *M*-estimators of multivariate location and scatter. *Ann. Statist.*, **1**, 51–67.
SMALL, N. J. H. (1978). Plotting squared radii. *Biometrika*, **65**, 657–658.
WILK, M. B., GNANADESIKAN, R. and HUYETT, M. J. (1962). Estimation of the parameters of the gamma distribution using order statistics. *Biometrika*, **49**, 525–545.