Chapter 11

CMCD Applications

11.1 DD Plots

A basic way of designing a graphical display is to arrange for reference situations to correspond to straight lines in the plot.
Chambers, Cleveland, Kleiner, and Tukey (1983, p. 322)

Definition 11.1: Rousseeuw and Van Driessen (1999). The *DD plot* is a plot of the classical Mahalanobis distances MD_i versus robust Mahalanobis distances RD_i.

The DD plot is analogous to the RR and FF plots and is used as a diagnostic for multivariate normality, elliptical symmetry and for outliers. Assume that the data set consists of iid vectors from an $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with second moments. Then the classical sample mean and covariance matrix $(T_M, \boldsymbol{C}_M) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ is a consistent estimator for $(\boldsymbol{\mu}, c_{\boldsymbol{X}}\boldsymbol{\Sigma}) = (E(\boldsymbol{X}), \text{Cov}(\boldsymbol{X}))$. Assume that an alternative algorithm estimator $(T_A, \boldsymbol{C}_A)$ is a consistent estimator for $(\boldsymbol{\mu}, a_A\boldsymbol{\Sigma})$ for some constant $a_A > 0$. By scaling the algorithm estimator, the DD plot can be constructed to follow the identity line with unit slope and zero intercept. Let $(T_R, \boldsymbol{C}_R) = (T_A, \boldsymbol{C}_A/\tau^2)$ denote the scaled algorithm estimator where $\tau > 0$ is a constant to be determined. Notice that $(T_R, \boldsymbol{C}_R)$ is a valid estimator of location and dispersion. Hence the robust distances used in the DD plot are given by

$$\text{RD}_i = \text{RD}_i(T_R, \boldsymbol{C}_R) = \sqrt{(\boldsymbol{x}_i - T_R(\boldsymbol{W}))^T[\boldsymbol{C}_R(\boldsymbol{W})]^{-1}(\boldsymbol{x}_i - T_R(\boldsymbol{W}))}$$

$= \tau \; D_i(T_A, \boldsymbol{C}_A)$ for $i = 1, ..., n$.

The following proposition shows that if consistent estimators are used to construct the distances, then the DD plot will tend to cluster tightly about the line segment through $(0,0)$ and $(\text{MD}_{n,\alpha}, \text{RD}_{n,\alpha})$ where $0 < \alpha < 1$ and $\text{MD}_{n,\alpha}$ is the α sample percentile of the MD_i. Nevertheless, the variability in the DD plot may increase with the distances. Let $K > 0$ be a constant, eg the 99th percentile of the χ^2_p distribution.

Proposition 11.1. Assume that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid observations from a distribution with parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a symmetric positive definite matrix. Let $a_j > 0$ and assume that $(\hat{\boldsymbol{\mu}}_{j,n}, \hat{\boldsymbol{\Sigma}}_{j,n})$ are consistent estimators of $(\boldsymbol{\mu}, a_j\boldsymbol{\Sigma})$ for $j = 1, 2$. Let $D_{i,j} \equiv D_i(\hat{\boldsymbol{\mu}}_{j,n}, \hat{\boldsymbol{\Sigma}}_{j,n})$ be the ith Mahalanobis distance computed from $(\hat{\boldsymbol{\mu}}_{j,n}, \hat{\boldsymbol{\Sigma}}_{j,n})$. Consider the cases in the region $R = \{i | 0 \leq D_{i,j} \leq K, \ j = 1, 2\}$. Let r_n denote the correlation between $D_{i,1}$ and $D_{i,2}$ for the cases in R (thus r_n is the correlation of the distances in the "lower left corner" of the DD plot). Then $r_n \to 1$ in probability as $n \to \infty$.

Proof. Let B_n denote the subset of the sample space on which both $\hat{\boldsymbol{\Sigma}}_{1,n}$ and $\hat{\boldsymbol{\Sigma}}_{2,n}$ have inverses. Then $P(B_n) \to 1$ as $n \to \infty$. The result follows if $D^2_j \xrightarrow{P} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})/a_j$ for fixed $\boldsymbol{x}$. This convergence holds since

$$D^2_j \equiv (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j} - \frac{\boldsymbol{\Sigma}^{-1}}{a_j} + \hat{\boldsymbol{\Sigma}}_j^{-1} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)$$

$$= (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{-\boldsymbol{\Sigma}^{-1}}{a_j} + \hat{\boldsymbol{\Sigma}}_j^{-1} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) + (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)$$

$$= \frac{1}{a_j} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T (-\boldsymbol{\Sigma}^{-1} + a_j \ \hat{\boldsymbol{\Sigma}}_j^{-1}) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) \ +$$

$$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j} \right) (\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)$$

$$= \frac{1}{a_j} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$$

$$+ \frac{2}{a_j} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j) + \frac{1}{a_j} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)$$

$$+ \frac{1}{a_j} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T [a_j \hat{\boldsymbol{\Sigma}}_j^{-1} - \boldsymbol{\Sigma}^{-1}] (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) \qquad (11.1)$$

on B_n, and the last three terms converge to zero in probability. QED

The above result implies that a plot of the MD_i versus the $D_i(T_A, \boldsymbol{C}_A) \equiv D_i(A)$ will follow a line through the origin with some positive slope since if $\boldsymbol{x} = \boldsymbol{\mu}$, then both the classical and the algorithm distances should be close to zero. We want to find τ such that $\text{RD}_i = \tau \ D_i(T_A, \boldsymbol{C}_A)$ and the DD plot of MD_i versus RD_i follows the identity line. By Proposition 11.1, the plot of MD_i versus $D_i(A)$ will follow the line segment defined by the origin $(0, 0)$ and the point of observed median Mahalanobis distances, $(\text{med}(\text{MD}_i), \text{med}(D_i(A)))$. This line segment has slope

$$\text{med}(D_i(A))/\text{med}(\text{MD}_i)$$

which is generally not one. By taking $\tau = \text{med}(\text{MD}_i)/\text{med}(D_i(A))$, the plot will follow the identity line if $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ is a consistent estimator of $(\boldsymbol{\mu}, c_{\boldsymbol{x}}\boldsymbol{\Sigma})$ and if $(T_A, \boldsymbol{C}_A)$ is a consistent estimator of $(\boldsymbol{\mu}, a_A\boldsymbol{\Sigma})$. (Using the notation from Proposition 11.1, let $(a_1, a_2) = (c_{\boldsymbol{x}}, a_A)$.) The classical estimator is consistent if the population has a nonsingular covariance matrix. The algorithm estimators $(T_A, \boldsymbol{C}_A)$ from Theorem 10.16 are consistent on the class of EC distributions that have a nonsingular covariance matrix, but are biased for non–EC distributions.

By replacing the observed median $\text{med}(\text{MD}_i)$ of the classical Mahalanobis distances with the target population analog, say MED, τ can be chosen so that the DD plot is *simultaneously* a diagnostic for elliptical symmetry and a diagnostic for the target EC distribution. That is, the plotted points follow the identity line if the data arise from a target EC distribution such as the multivariate normal distribution, but the points follow a line with non-unit slope if the data arise from an alternative EC distribution. In addition the DD plot can often detect departures from elliptical symmetry such as outliers, the presence of two groups, or the presence of a mixture distribution. These facts make the DD plot a useful alternative to other graphical diagnostics for target distributions. See Easton and McCulloch (1990), Li, Fang, and Zhu (1997), and Liu, Parelius, and Singh (1999) for references.

Example 11.1. Rousseeuw and Van Driessen (1999) choose the multivariate normal $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution as the target. If the data are indeed iid MVN vectors, then the $(\text{MD}_i)^2$ are asymptotically χ^2_p random variables, and $\text{MED} = \sqrt{\chi^2_{p,0.5}}$ where $\chi^2_{p,0.5}$ is the median of the χ^2_p distribution. Since the target distribution is Gaussian, let

$$\mathrm{RD}_i = \frac{\sqrt{\chi^2_{p,0.5}}}{\mathrm{med}(D_i(A))} D_i(A) \quad \text{so that} \quad \tau = \frac{\sqrt{\chi^2_{\mathrm{p},0.5}}}{\mathrm{med}(\mathrm{D_i(A)})}. \tag{11.2}$$

Note that the DD plot can be tailored to follow the identity line if the data are iid observations from any target elliptically contoured distribution that has nonsingular covariance matrix. If it is known that med(MD$_i$) $\approx$ MED where MED is the target population analog (obtained, for example, via simulation, or from the actual target distribution as in Equations (10.8), (10.9) and (10.10) on p. 308), then use

$$\mathrm{RD}_i = \tau\ D_i(A) = \frac{\mathrm{MED}}{\mathrm{med}(D_i(A))} D_i(A). \tag{11.3}$$

The choice of the algorithm estimator $(T_A, \boldsymbol{C}_A)$ is important, and the HB $\sqrt{n}$ consistent FCH estimator is a good choice. In this chapter we used the *R/Splus* function `cov.mcd` which is basically an implementation of the elemental MCD concentration algorithm described in the previous chapter. The number of starts used was $K = \max(500, n/10)$ (the default is $K = 500$, so the default can be used if $n \leq 5000$).

Conjecture 11.1. If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ and an elemental MCD concentration algorithm is used to produce the estimator $(T_{A,n}, \boldsymbol{C}_{A,n})$, then this algorithm estimator is consistent for $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ for some constant $a > 0$ (that depends on g) if the number of starts $K = K(n) \to \infty$ as the sample size $n \to \infty$.

Notice that if this conjecture is true, and if the data is EC with 2nd moments, then

$$\left[\frac{\mathrm{med}(D_i(A))}{\mathrm{med}(\mathrm{MD}_i)}\right]^2 \boldsymbol{C}_A \tag{11.4}$$

estimates Cov($\boldsymbol{X}$). For the DD plot, consistency is desirable but not necessary. It is necessary that the correlation of the smallest 99% of the MD$_i$ and RD$_i$ be very high. This correlation goes to 1 by Proposition 11.1 if consistent estimators are used.

The choice of using a concentration algorithm to produce $(T_A, \boldsymbol{C}_A)$ is certainly not perfect, and the `cov.mcd` estimator should be modified by adding

Table 11.1: **Corr(RD_i, MD_i) for $N_p(\mathbf{0}, \boldsymbol{I}_p)$ Data, 100 Runs.**

p	n	mean	min	% < 0.95	% < 0.8
3	44	0.866	0.541	81	20
3	100	0.967	0.908	24	0
7	76	0.843	0.622	97	26
10	100	0.866	0.481	98	12
15	140	0.874	0.675	100	6
15	200	0.945	0.870	41	0
20	180	0.889	0.777	100	2
20	1000	0.998	0.996	0	0
50	420	0.894	0.846	100	0

the FCH starts as shown in Theorem 10.17. There exist data sets with outliers or two groups such that both the classical and robust estimators produce ellipsoids that are nearly concentric. We suspect that the situation worsens as p increases.

In a simulation study, $N_p(\mathbf{0}, \boldsymbol{I}_p)$ data were generated and `cov.mcd` was used to compute first the $D_i(A)$, and then the RD$_i$ using Equation (11.2). The results are shown in Table 11.1. Each choice of n and p used 100 runs, and the 100 correlations between the RD$_i$ and the MD$_i$ were computed. The mean and minimum of these correlations are reported along with the percentage of correlations that were less than 0.95 and 0.80. The simulation shows that small data sets (of roughly size $n < 8p + 20$) yield plotted points that may not cluster tightly about the identity line even if the data distribution is Gaussian.

Since every estimator of location and dispersion defines an ellipsoid, the DD plot can be used to examine which points are in the robust ellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - T_R)^T \boldsymbol{C}_R^{-1} (\boldsymbol{x} - T_R) \le RD^2_{(h)}\} \tag{11.5}$$

where $RD^2_{(h)}$ is the hth smallest squared robust Mahalanobis distance, and which points are in a classical ellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x} - \overline{\boldsymbol{x}}) \le MD^2_{(h)}\}. \tag{11.6}$$

In the DD plot, points below $RD_{(h)}$ correspond to cases that are in the ellipsoid given by Equation (11.5) while points to the left of $MD_{(h)}$ are in an ellipsoid determined by Equation (11.6).

The DD plot will follow a line through the origin closely if the two ellipsoids are nearly concentric, eg if the data is EC. The DD plot will follow the identity line closely if med(MD$_i$) $\approx$ MED, and RD$_i^2$ =

$$(\boldsymbol{x}_i - T_A)^T[(\frac{\text{MED}}{\text{med}(D_i(A))})^2 \boldsymbol{C}_A^{-1}](\boldsymbol{x}_i - T_A) \approx (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}) = \text{MD}_i^2$$

for $i = 1, ..., n$. When the distribution is not EC,

$$(T_A, \boldsymbol{C}_A) = (T_{FCH}, \boldsymbol{C}_{FCH}) \;\text{ or }\; (\text{T}_\text{A}, \boldsymbol{C}_\text{A}) = (\text{T}_\text{FMCD}, \boldsymbol{C}_\text{FMCD})$$

and $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ will often produce ellipsoids that are far from concentric.

Application 11.1. The DD plot can be used *simultaneously* as a diagnostic for whether the data arise from a multivariate normal (MVN or Gaussian) distribution or from another EC distribution with nonsingular covariance matrix. EC data will cluster about a straight line through the origin; MVN data in particular will cluster about the identity line. Thus the DD plot can be used to assess the success of numerical transformations towards elliptical symmetry. This application is important since many statistical methods assume that the underlying data distribution is MVN or EC.

For this application, the RFCH estimator may be best. For MVN data, the RD$_i$ from the RFCH estimator tend to have a higher correlation with the MD$_i$ from the classical estimator than the RD$_i$ from the FCH estimator, and the `cov.mcd` estimator may be inconsistent.

Figure 11.1 shows the DD plots for 3 artificial data sets using `cov.mcd`. The DD plot for 200 $N_3(\boldsymbol{0}, \boldsymbol{I}_3)$ points shown in Figure 1a resembles the identity line. The DD plot for 200 points from the elliptically contoured distribution $0.6N_3(\boldsymbol{0}, \boldsymbol{I}_3) + 0.4N_3(\boldsymbol{0}, 25\ \boldsymbol{I}_3)$ in Figure 11.1b clusters about a line through the origin with a slope close to 2.0.

A *weighted DD plot* magnifies the lower left corner of the DD plot by omitting the cases with RD$_i \geq \sqrt{\chi^2_{p,.975}}$. This technique can magnify features that are obscured when large RD$_i$'s are present. If the distribution of $\boldsymbol{x}$ is EC with nonsingular $\boldsymbol{\Sigma}$, Proposition 11.1 implies that the correlation of the

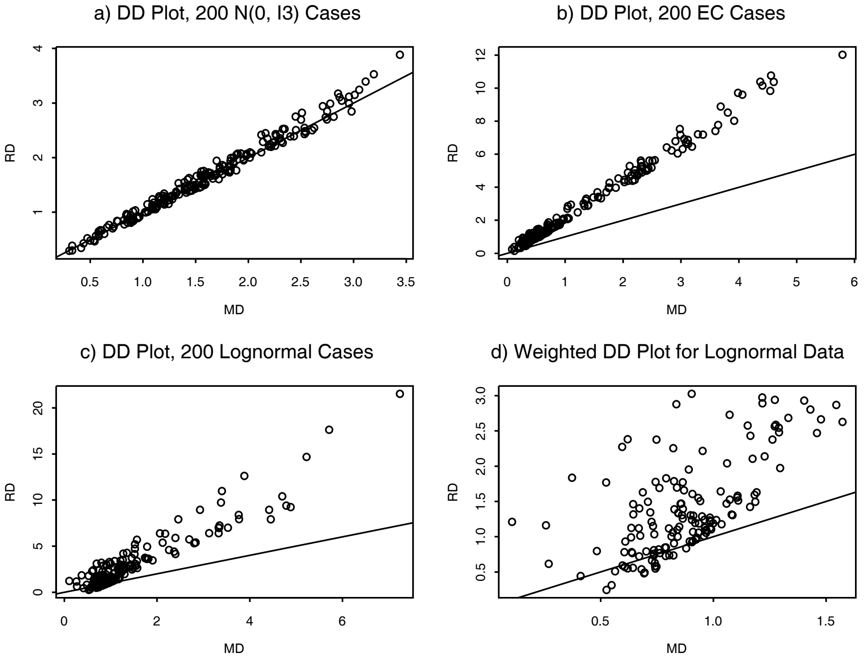

Figure 11.1: 4 DD Plots

points in the weighted DD plot will tend to one and that the points will cluster about a line passing through the origin. For example, the plotted points in the weighted DD plot (not shown) for the non-MVN EC data of Figure 11.1b are highly correlated and still follow a line through the origin with a slope close to 2.0.

Figures 11.1c and 11.1d illustrate how to use the weighted DD plot. The ith case in Figure 11.1c is $(\exp(x_{i,1}), \exp(x_{i,2}), \exp(x_{i,3}))^T$ where $\boldsymbol{x}_i$ is the ith case in Figure 11a; ie, the marginals follow a lognormal distribution. The plot does not resemble the identity line, correctly suggesting that the distribution of the data is not MVN; however, the correlation of the plotted points is rather high. Figure 11.1d is the weighted DD plot where cases with $\text{RD}_i \geq \sqrt{\chi^2_{3,.975}} \approx 3.06$ have been removed. Notice that the correlation of the plotted points is not close to one and that the best fitting line in Figure 11.1d may not pass through the origin. These results suggest that the distribution of $\boldsymbol{x}$ is not EC.

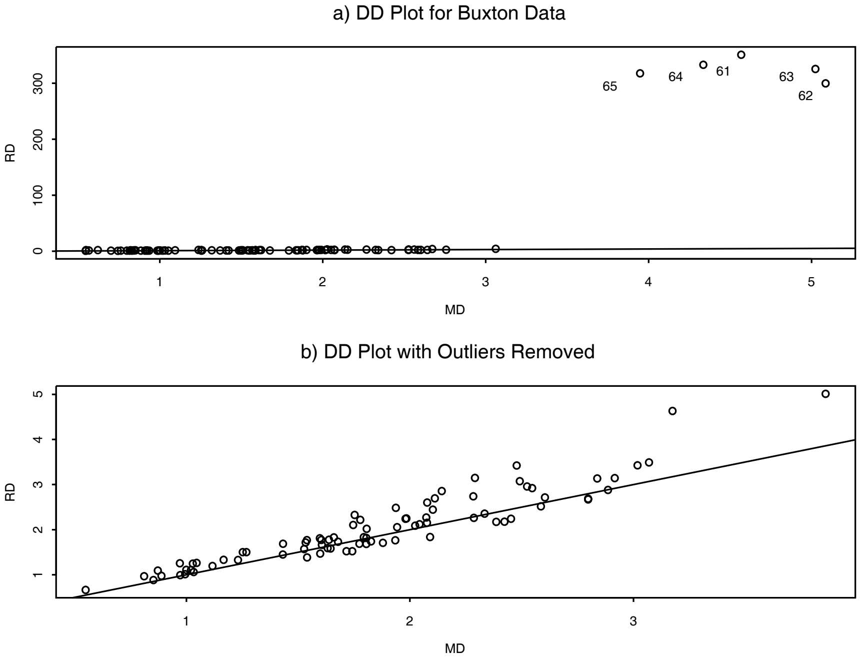

Figure 11.2: DD Plots for the Buxton Data

It is easier to use the DD plot as a diagnostic for a target distribution such as the MVN distribution than as a diagnostic for elliptical symmetry. If the data arise from the target distribution, then the DD plot will tend to be a useful diagnostic when the sample size n is such that the sample correlation coefficient in the DD plot is at least 0.80 with high probability. As a diagnostic for elliptical symmetry, it may be useful to add the OLS line to the DD plot and weighted DD plot as a visual aid, along with numerical quantities such as the OLS slope and the correlation of the plotted points.

Numerical methods for transforming data towards a target EC distribution have been developed. Generalizations of the Box–Cox transformation towards a multivariate normal distribution are described in Velilla (1993). Alternatively, Cook and Nachtsheim (1994) offer a two-step numerical procedure for transforming data towards a target EC distribution. The first step simply gives zero weight to a fixed percentage of cases that have the largest robust Mahalanobis distances, and the second step uses Monte Carlo case reweighting with Voronoi weights.

Example 11.2. Buxton (1920, p. 232-5) gives 20 measurements of 88 men. We will examine whether the multivariate normal distribution is a plausible model for the measurements *head length, nasal height, bigonal breadth,* and *cephalic index* where one case has been deleted due to missing values. Figure 11.2a shows the DD plot. Five head lengths were recorded to be around 5 feet and are massive outliers. Figure 11.2b is the DD plot computed after deleting these points and suggests that the normal distribution is plausible. (The recomputation of the DD plot means that the plot is not a weighted DD plot which would simply omit the outliers and then rescale the vertical axis.)

The DD plot complements rather than replaces the numerical procedures. For example, if the goal of the transformation is to achieve a multivariate normal distribution and if the data points cluster tightly about the identity line, as in Figure 11.1a, then perhaps no transformation is needed. For the data in Figure 11.1c, a good numerical procedure should suggest coordinatewise log transforms. Following this transformation, the resulting plot shown in Figure 11.1a indicates that the transformation to normality was successful.

Application 11.2. The DD plot can be used to detect multivariate outliers. See Figures 10.2 and 11.2a.

11.2 Robust Prediction Regions

Suppose that $(T_A, \boldsymbol{C}_A)$ denotes the algorithm estimator of location and dispersion. Section 11.1 showed that if $\boldsymbol{X}$ is multivariate normal $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, T_A estimates $\boldsymbol{\mu}$ and $\boldsymbol{C}_A/\tau^2$ estimates $\boldsymbol{\Sigma}$ where τ is given in Equation (11.2). Then $(T_R, \boldsymbol{C}_R) \equiv (T_A, \boldsymbol{C}_A/\tau^2)$ is an estimator of multivariate location and dispersion. Given an estimator $(T, \boldsymbol{C})$, a 95% *covering ellipsoid* for MVN data is the ellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \le \chi^2_{p,0.95}\}. \tag{11.7}$$

This ellipsoid is a large sample 95% prediction region if the data is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and if $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

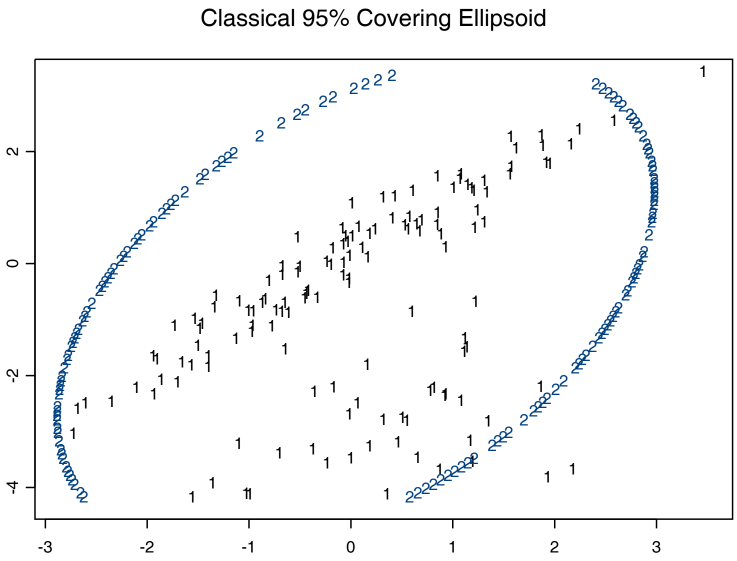

Figure 11.3: Artificial Bivariate Data

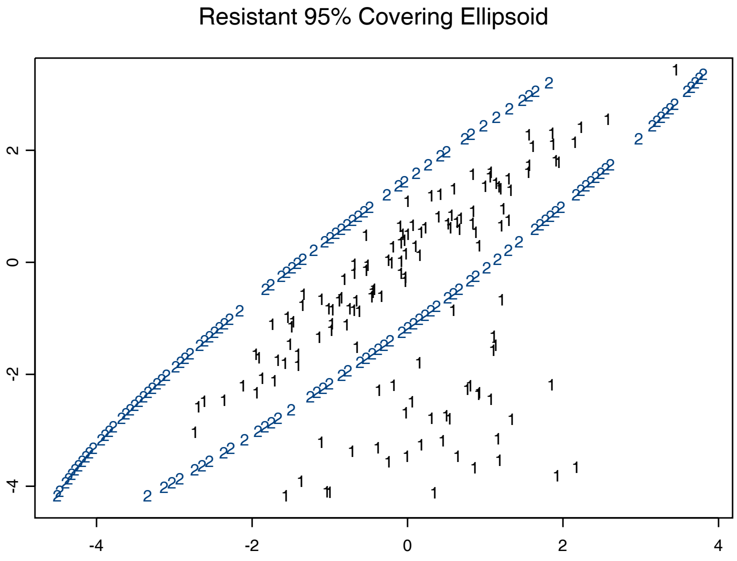

Figure 11.4: Artificial Data

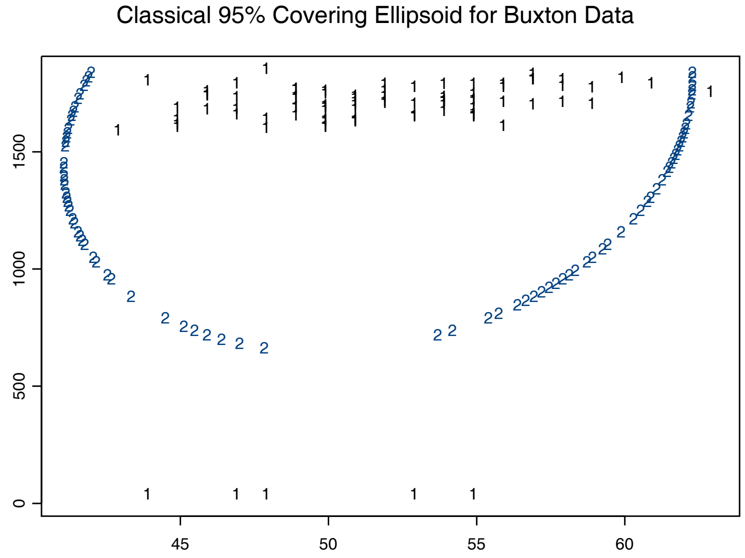

Figure 11.5: Ellipsoid is Inflated by Outliers

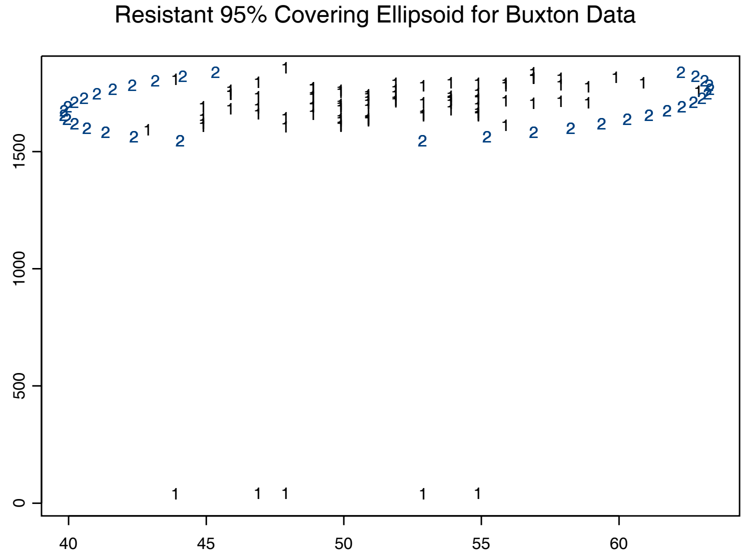

Figure 11.6: Ellipsoid Ignores Outliers

Example 11.3. An artificial data set consisting of 100 iid cases from a

$$N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1.49 & 1.4 \\ 1.4 & 1.49 \end{pmatrix} \right)$$

distribution and 40 iid cases from a bivariate normal distribution with mean $(0, -3)^T$ and covariance $\boldsymbol{I}_2$. Figure 11.3 shows the classical covering ellipsoid that uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. The symbol "1" denotes the data while the symbol "2" is on the border of the covering ellipse. Notice that the classical ellipsoid covers almost all of the data. Figure 11.4 displays the resistant covering ellipse. The resistant covering ellipse contains most of the 100 "clean" cases and excludes the 40 outliers. Problem 11.5 recreates similar figures with the classical and the resistant *R/Splus* `cov.mcd` estimators.

Example 11.4. Buxton (1920) gives various measurements on 88 men including *height* and *nasal height.* Five *heights* were recorded to be about 19mm and are massive outliers. Figure 11.5 shows that the classical covering ellipsoid is quite large but does not include any of the outliers. Figure 11.6 shows that the resistant covering ellipsoid is not inflated by the outliers.

11.3 Resistant Regression

Ellipsoidal trimming can be used to create resistant multiple linear regression (MLR) estimators. To perform ellipsoidal trimming, an estimator $(T, \boldsymbol{C})$ is computed and used to create the squared Mahalanobis distances D_i^2 for each vector of observed predictors $\boldsymbol{x}_i$. If the ordered distance $D_{(j)}$ is unique, then j of the $\boldsymbol{x}_i$'s are in the ellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) \le D^2_{(j)}\}. \tag{11.8}$$

The ith case $(y_i, \boldsymbol{x}_i^T)^T$ is trimmed if $D_i > D_{(j)}$. Then an estimator of $\boldsymbol{\beta}$ is computed from the remaining cases. For example, if $j \approx 0.9n$, then about 10% of the cases are trimmed, and OLS or L_1 could be used on the cases that remain.

Recall that a response plot is a plot of the fitted values $\hat{Y}_i$ versus the response Y_i and is very useful for detecting outliers. If the MLR model holds and the MLR estimator is good, then the plotted points will scatter about the identity line that has unit slope and zero intercept. The identity line is added to the plot as a visual aid, and the vertical deviations from the identity line are equal to the residuals since $Y_i - \hat{Y}_i = r_i$.

The resistant trimmed views estimator combines ellipsoidal trimming and the response plot. First compute $(T, \boldsymbol{C})$, perhaps using the FCH estimator or the *R/Splus* function `cov.mcd`. Trim the M% of the cases with the largest Mahalanobis distances, and then compute the MLR estimator $\hat{\boldsymbol{\beta}}_M$ from the remaining cases. Use $M =$ 0, 10, 20, 30, 40, 50, 60, 70, 80, and 90 to generate ten response plots of the fitted values $\hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}_i$ versus y_i using all n cases. (Fewer plots are used for small data sets if $\hat{\boldsymbol{\beta}}_M$ can not be computed for large M.) These plots are called "trimmed views."

Definition 11.2. The trimmed views (TV) estimator $\hat{\boldsymbol{\beta}}_{T,n}$ corresponds to the trimmed view where the bulk of the plotted points follow the identity line with smallest variance function, ignoring any outliers.

Example 11.4 (continued). For the Buxton (1920) data, *height* was the response variable while an intercept, *head length, nasal height, bigonal breadth,* and *cephalic index* were used as predictors in the multiple linear regression model. Observation 9 was deleted since it had missing values. Five individuals, cases 61–65, were reported to be about 0.75 inches tall with head lengths well over five feet! OLS was used on the cases remaining after trimming, and Figure 11.7 shows four trimmed views corresponding to 90%, 70%, 40% and 0% trimming. The OLS TV estimator used 70% trimming since this trimmed view was best. Since the vertical distance from a plotted point to the identity line is equal to the case's residual, the outliers had massive residuals for 90%, 70% and 40% trimming. Notice that the OLS trimmed view with 0% trimming "passed through the outliers" since the cluster of outliers is scattered about the identity line.

The TV estimator $\hat{\boldsymbol{\beta}}_{T,n}$ has good statistical properties if an estimator with good statistical properties is applied to the cases $(\boldsymbol{X}_{M,n}, \boldsymbol{Y}_{M,n})$ that remain after trimming. Candidates include OLS, L_1, Huber's M–estimator, Mallows' GM–estimator or the Wilcoxon rank estimator. See Rousseeuw and Leroy (1987, p. 12-13, 150). The basic idea is that if an estimator with $O_P(n^{-1/2})$ convergence rate is applied to a set of $n_M \propto n$ cases, then the resulting estimator $\hat{\boldsymbol{\beta}}_{M,n}$ also has $O_P(n^{-1/2})$ rate provided that the response y was not used to select the n_M cases in the set. If $\|\hat{\boldsymbol{\beta}}_{M,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$ for $M = 0, ..., 90$ then $\|\hat{\boldsymbol{\beta}}_{T,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$ by Pratt (1959).

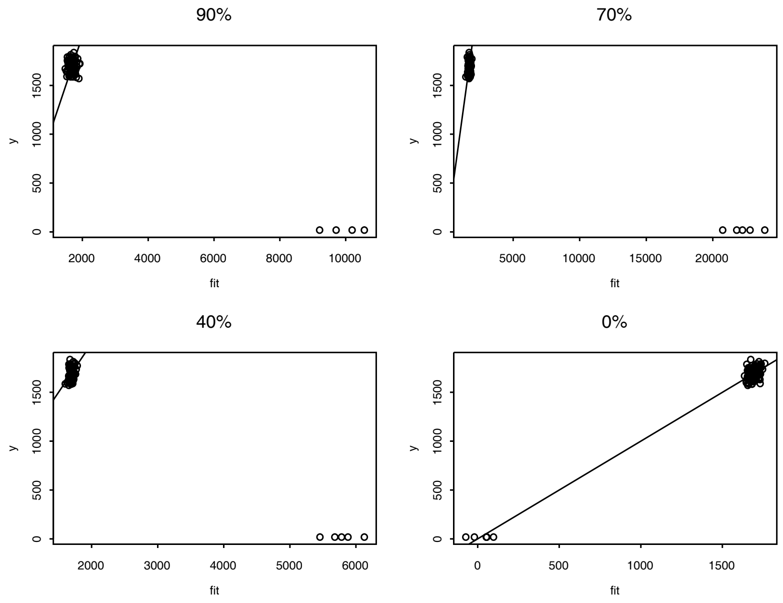

Figure 11.7: 4 Trimmed Views for the Buxton Data

Let $\boldsymbol{X}_n = \boldsymbol{X}_{0,n}$ denote the full design matrix. Often when proving asymptotic normality of an MLR estimator $\hat{\boldsymbol{\beta}}_{0,n}$, it is assumed that

$$\frac{\boldsymbol{X}_n^T \boldsymbol{X}_n}{n} \to \boldsymbol{W}^{-1}.$$

If $\hat{\boldsymbol{\beta}}_{0,n}$ has $O_P(n^{-1/2})$ rate and if for big enough n all of the diagonal elements of

$$\left(\frac{\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n}}{n} \right)^{-1}$$

are all contained in an interval $[0, B)$ for some $B > 0$, then $\|\hat{\boldsymbol{\beta}}_{M,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$.

The distribution of the estimator $\hat{\boldsymbol{\beta}}_{M,n}$ is especially simple when OLS is used and the errors are iid $N(0, \sigma^2)$. Then

$$\hat{\boldsymbol{\beta}}_{M,n} = (\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n})^{-1} \boldsymbol{X}_{M,n}^T \boldsymbol{Y}_{M,n} \sim N_p(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n})^{-1})$$

and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{M,n} - \boldsymbol{\beta}) \sim N_p(\boldsymbol{0}, \sigma^2 (\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n}/n)^{-1})$. Notice that this result does not imply that the distribution of $\hat{\boldsymbol{\beta}}_{T,n}$ is normal.

Table 11.2: Summaries for Seven Data Sets, the Correlations of the Residuals from TV(M) and the Alternative Method are Given in the 1st 5 Rows

Method	Buxton	Gladstone	glado	hbk	major	nasty	wood
MBA	0.997	1.0	0.455	0.960	1.0	-0.004	0.9997
LMSREG	-0.114	0.671	0.938	0.977	0.981	0.9999	0.9995
LTSREG	-0.048	0.973	0.468	0.272	0.941	0.028	0.214
L1	-0.016	0.983	0.459	0.316	0.979	0.007	0.178
OLS	0.011	1.0	0.459	0.780	1.0	0.009	0.227
outliers	61-65	none	119	1-10	3,44	2,6,...,30	4,6,8,19
n	87	247	247	75	112	32	20
p	5	7	7	4	6	5	6
M	70	0	30	90	0	90	20

Table 11.2 compares the TV, MBA (for MLR), `lmsreg`, `ltsreg`, L_1 and OLS estimators on 7 data sets available from the text's website. The column headers give the file name while the remaining rows of the table give the sample size n, the number of predictors p, the amount of trimming M used by the TV estimator, the correlation of the residuals from the TV estimator with the corresponding alternative estimator, and the cases that were outliers. If the correlation was greater than 0.9, then the method was effective in detecting the outliers, and the method failed, otherwise. Sometimes the trimming percentage M for the TV estimator was picked after fitting the bulk of the data in order to find the good leverage points and outliers.

Notice that the TV, MBA and OLS estimators were the same for the Gladstone data and for the *major* data (Tremearne 1911) which had two small y–outliers. For the Gladstone data, there is a cluster of infants that are good leverage points, and we attempt to predict *brain weight* with the head measurements *height, length, breadth, size* and *cephalic index.* Originally, the variable *length* was incorrectly entered as 109 instead of 199 for case 119, and the *glado* data contains this outlier. In 1997, `lmsreg` was not able to detect the outlier while `ltsreg` did. Due to changes in the *Splus* 2000 code, `lmsreg` now detects the outlier but `ltsreg` does not.

The TV estimator can be modified to create a resistant weighted MLR estimator. To see this, recall that the weighted least squares (WLS) estimator using weights W_i can be found using the ordinary least squares (OLS) regression (without intercept) of $\sqrt{W_i}Y_i$ on $\sqrt{W_i}\boldsymbol{x}_i$. This idea can be used for categorical data analysis since the minimum chi-square estimator is often computed using WLS. See Section 13.4 for an illustration of Application 11.3 below. Let $\boldsymbol{x}_i = (1, x_{i,2}, ..., x_{i,p})^T$, let $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ and let $\tilde{\boldsymbol{\beta}}$ be an estimator of $\boldsymbol{\beta}$.

Definition 11.3. For a multiple linear regression model with weights W_i, a **weighted response plot** is a plot of $\sqrt{W_i}\boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}$ versus $\sqrt{W_i}Y_i$. The **weighted residual plot** is a plot of $\sqrt{W_i}\boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}$ versus the WMLR residuals $r_{Wi} = \sqrt{W_i}Y_i - \sqrt{W_i}\boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}$.

Application 11.3. For resistant weighted MLR, use the WTV estimator which is selected from ten weighted response plots.

11.4 Robustifying Robust Estimators

Many papers have been written that need a HB consistent estimator of MLD. Since no practical HB estimator was available, inconsistent zero breakdown estimators were often used in implementations, resulting in zero breakdown estimators that were often inconsistent (although perhaps useful as diagnostics).

Applications of the robust $\sqrt{n}$ consistent CMCD and FCH estimators are numerous. For example, robustify the ideas in the following papers by using the FCH estimator instead of the FMCD, MCD or MVE estimator. *Binary regression:* see Croux and Haesbroeck (2003). *Canonical correlation analysis:* see Branco, Croux, Filzmoser, and Oliviera (2005). *Discriminant analysis:* see Hubert and Van Driessen (2004). *Factor analysis:* see Pison, Rousseeuw, Filzmoser, and Croux (2003). *Generalized partial linear models:* see He, Fung and Zhu (2005). *Analogs of Hotelling's T^2 test:* see Willems, Pison, Rousseeuw, and Van Aelst (2002). *Longitudinal data analysis:* see He, Cui and Simpson (2004). *Multivariate analysis diagnostics:* the DD plot of classical Mahalanobis distances versus FCH distances should be used for multivariate analysis much as Cook's distances are used for MLR. *Multivariate regression:* see Agulló, Croux and Van Aelst (2008). *Principal components:* see Hubert, Rousseeuw, and Vanden Branden (2005) and Croux, Filzmoser, and Oliveira (2007). *Efficient estimators of MLD:* see He and Wang (1996).

Also see Hubert, Rousseeuw and Van Aelst (2008) for references. Their FMCD and FLTS estimators do not compute the MCD and LTS estimators, and need to be modified as in Remarks 8.8 and 10.5.

Regression via Dimension Reduction: Regression is the study of the conditional distribution of the response Y given the vector of predictors $\boldsymbol{x} = (1, \boldsymbol{w}^T)^T$ where $\boldsymbol{w}$ is the vector of nontrivial predictors. Make a DD plot of the classical Mahalanobis distances versus the robust distances computed from $\boldsymbol{w}$. If $\boldsymbol{w}$ comes from an elliptically contoured distribution, then the plotted points in the DD plot should follow a straight line through the origin. Give zero weight to cases in the DD plot that do not cluster tightly about "the best straight line" through the origin (often the identity line with unit slope), and run a weighted regression procedure. This technique can increase the resistance of regression procedures such as sliced inverse regression (SIR, see Li, 1991) and MAVE (Xia, Tong, Li, and Zhu, 2002). Also see Chang and Olive (2007), Cook and Nachtsheim (1994) and Li, Cook and Nachtsheim (2004).

Visualizing 1D Regression: A 1D regression is a special case of regression where the response Y is independent of the predictors $\boldsymbol{x}$ given $\boldsymbol{\beta}^T\boldsymbol{x}$. Generalized linear models and single index models are important special cases. Resistant methods for visualizing 1D regression are given in Olive (2002, 2004b). Also see Chapters 12 and 13.

11.5 Complements

The first section of this chapter followed Olive (2002) closely. The DD plot can be used to diagnose elliptical symmetry, to detect outliers, and to assess the success of numerical methods for transforming data towards an elliptically contoured distribution. Since many statistical methods assume that the underlying data distribution is Gaussian or EC, there is an enormous literature on numerical tests for elliptical symmetry. Bogdan (1999), Czörgö (1986) and Thode (2002) provide references for tests for multivariate normality while Koltchinskii and Li (1998) and Manzotti, Pérez and Quiroz (2002) have references for tests for elliptically contoured distributions.

The TV estimator was proposed by Olive (2002, 2005) and is similar to an estimator proposed by Rousseeuw and van Zomeren (1992). Although both the TV and MBA estimators have the good $O_P(n^{-1/2})$ convergence rate, their efficiency under normality may be very low. Chang and Olive (2008) suggest a method of adaptive trimming such that the resulting estimator is asymptotically equivalent to the OLS estimator. Also see Section 12.5. High breakdown estimators that have high efficiency tend to be impractical to compute, but exceptions include the estimators from Theorem 8.8 and Remark 8.7.

The ideas used in Section 11.3 have the potential for making many methods resistant. First, suppose that the MLR model holds but $\text{Cov}(\boldsymbol{e}) = \sigma^2 \boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{V}'$ where $\boldsymbol{V}$ is known and nonsingular. Then $\boldsymbol{V}^{-1}\boldsymbol{Y} = \boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{V}^{-1}\boldsymbol{e}$, and the TV and MBA MLR estimators can be applied to $\tilde{\boldsymbol{Y}} = \boldsymbol{V}^{-1}\boldsymbol{Y}$ and $\tilde{\boldsymbol{X}} = \boldsymbol{V}^{-1}\boldsymbol{X}$ provided that OLS is fit without an intercept.

Secondly, many 1D regression models (where Y_i is independent of $\boldsymbol{x}_i$ given the sufficient predictor $\boldsymbol{x}_i^T\boldsymbol{\beta}$) can be made resistant by making EY plots of the estimated sufficient predictor $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_M$ versus Y_i for the 10 trimming proportions. Since 1D regression is the study of the conditional distribution of Y_i given $\boldsymbol{x}_i^T\boldsymbol{\beta}$, the EY plot is used to visualize this distribution and needs to be made anyway. See Chapter 12.

Thirdly, for nonlinear regression models of the form $Y_i = m(\boldsymbol{x}_i, \boldsymbol{\beta}) + e_i$, the fitted values are $\hat{Y}_i = m(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}})$ and the residuals are $r_i = Y_i - \hat{Y}_i$. The points in the FY plot of the fitted values versus the response should follow the identity line. The TV estimator would make FY and residual plots for each of the trimming proportions. The MBA estimator with the median squared residual criterion can also be used for many of these models.

Mϕller, von Frese and Bro (2005) is a good illustration of the widespread use of inconsistent zero breakdown estimators plugged in place of classical estimators in an attempt to make the multivariate method robust.

11.6 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

11.1*. If X and Y are random variables, show that

$$\text{Cov(X, Y)} = [\text{Var(X + Y)} - \text{Var(X - Y)}]/4.$$

R/Splus Problems

Warning: Use the command *source("A:/rpack.txt")* **to download the programs. See Preface or Section 14.2.** Typing the name of the `rpack` function, eg *ddplot*, will display the code for the function. Use the `args` command, eg *args(ddplot)*, to display the needed arguments for the function.

11.2. a) Download the program `ddsim`. (In *R*, type the command *library(MASS)*.)

b) Using the function *ddsim* for $p = 2, 3, 4$, determine how large the sample size n should be in order for the DD plot of n $N_p(\mathbf{0}, \boldsymbol{I}_p)$ cases to be cluster tightly about the identity line with high probability. Table your results. (Hint: type the command *ddsim(n=20,p=2)* and increase n by 10 until most of the 20 plots look linear. Then repeat for $p = 3$ with the n that worked for $p = 2$. Then repeat for $p = 4$ with the n that worked for $p = 3$.)

11.3. a) Download the program `corrsim`. (In *R*, type the command *library(MASS)*.)

b) A numerical quantity of interest is the correlation between the MD_i and RD_i in a DD plot that uses n $N_p(\mathbf{0}, \boldsymbol{I}_p)$ cases. Using the function *corrsim* for $p = 2, 3, 4$, determine how large the sample size n should be in order for 9 out of 10 correlations to be greater than 0.9. (Try to make n small.) Table your results. (Hint: type the command *corrsim(n=20,p=2,nruns=10)* and increase n by 10 until 9 or 10 of the correlations are greater than 0.9. Then repeat for $p = 3$ with the n that worked for $p = 2$. Then repeat for $p = 4$ with the n that worked for $p = 3$.)

11.4***.** a) Download the `ddplot` function. (In *R*, type the command *library(MASS)*.)

b) Using the following commands to make generate data from the EC distribution $(1 - \epsilon)N_p(\mathbf{0}, \boldsymbol{I}_p) + \epsilon N_p(\mathbf{0}, 25\ \boldsymbol{I}_p)$ where $p = 3$ and $\epsilon = 0.4$.

```
n <- 400
p <- 3
eps <- 0.4
x <- matrix(rnorm(n * p), ncol = p, nrow = n)
zu <- runif(n)
x[zu < eps,] <- x[zu < eps,]*5
```

c) Use the command `ddplot(x)` to make a DD plot and include the plot in *Word.* What is the slope of the line followed by the plotted points?

11.5. a) Download the `ellipse` function.

b) Use the following commands to create a bivariate data set with outliers and to obtain a classical and robust covering ellipsoid. Include the two plots in *Word.* (In *R*, type the command *library(MASS).*)

```
> simx2 <- matrix(rnorm(200),nrow=100,ncol=2)
> outx2 <- matrix(10 + rnorm(80),nrow=40,ncol=2)
> outx2 <- rbind(outx2,simx2)
> ellipse(outx2)
> zout <- cov.mcd(outx2)
> ellipse(outx2,center=zout$center,cov=zout$cov)
```

11.6. a) Download the function `mplot`.

b) Enter the commands in Problem 11.4b to obtain a data set x. The function `mplot` makes a plot without the RD_i and the slope of the resulting line is of interest.

c) Use the command `mplot(x)` and place the resulting plot in *Word.*

d) Do you prefer the DD plot or the mplot? Explain.

11.7 a) Download the function `wddplot`.

b) Enter the commands in Problem 11.4b to obtain a data set x.

c) Use the command `wddplot(x)` and place the resulting plot in *Word.*

11.8. a) In addition to the *source("A:/rpack.txt")* command, also use the *source("A:/robdata.txt")* command (and in *R*, type the *library(MASS)* command).

b) Type the command *tvreg(buxx,buxy,ii=1).* Click the rightmost mouse button (and in *R*, highlight *Stop*). The forward response plot should appear. Repeat 10 times and remember which plot percentage M (say M = 0) had the best forward response plot. Then type the command *tvreg2(buxx,buxy, M = 0)* (except use your value of M, not 0). Again, click the rightmost mouse button (and in *R*, highlight *Stop*). The forward response plot should appear. Hold down the *Ctrl* and *c* keys to make a copy of the plot. Then paste the plot in *Word.*

c) The estimated coefficients $\hat{\boldsymbol{\beta}}_{TV}$ from the best plot should have appeared on the screen. Copy and paste these coefficients into *Word.*