Chapter 5

DD Plots and Prediction Regions

5.1 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 5.1: 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 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 $(\boldsymbol{\mu}, c_{\boldsymbol{X}}\boldsymbol{\Sigma}) = (E(\boldsymbol{X}), \mathrm{Cov}(\boldsymbol{X}))$. 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

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

$= \tau\ D_i(T_A, \boldsymbol{C}_A)$ 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 $(\mathrm{MD}_{n,\alpha}, \mathrm{RD}_{n,\alpha})$ where $0 < \alpha < 1$ and $\mathrm{MD}_{n,\alpha}$ 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 5.1. 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 $j = 1, 2$.

a) $D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) - \frac{1}{a_j} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = o_P(1)$.

b) Let $0 < \delta \leq 0.5$. If $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) - (\boldsymbol{\mu}, a_j\boldsymbol{\Sigma}) = O_p(n^{-\delta})$ and $a_j\hat{\boldsymbol{\Sigma}}_j^{-1} - \boldsymbol{\Sigma}^{-1} = O_P(n^{-\delta})$, then

$$D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) - \frac{1}{a_j} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = O_P(n^{-\delta}).$$

c) 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 $n \to \infty$.

a) and b): $D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) =$

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

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

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

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

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

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

$$+\frac{1}{a_j}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_j)^T[a_j\hat{\boldsymbol{\Sigma}}_j^{-1}-\boldsymbol{\Sigma}^{-1}](\boldsymbol{x}-\hat{\boldsymbol{\mu}}_j) \quad (5.1)$$

on B_n, and the last three terms are $o_P(1)$ under a) and $O_P(n^{-\delta})$ under b).

c) Following the proof of a), $D_j^2 \equiv D_{\boldsymbol{x}}^2(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) \xrightarrow{P} (\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/a_j$ for fixed $\boldsymbol{x}$, and the result follows.

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 $\mathrm{RD}_i = \tau\ D_i(T_A, \boldsymbol{C}_A)$ and the DD plot of MD_i versus RD_i follows the identity line. By Proposition 5.1, 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, $(\mathrm{med}(\mathrm{MD}_i), \mathrm{med}(D_i(A)))$. This line segment has slope

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

which is generally not one. By taking $\tau = \mathrm{med}(\mathrm{MD}_i)/\mathrm{med}(D_i(A))$, 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 5.1, 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 4.10 are consistent on a large class of EC distributions that have a nonsingular covariance matrix, but tend to be biased for non–EC distributions.

By replacing the observed median $\mathrm{med}(\mathrm{MD}_i)$ 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 5.1. 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 $(\mathrm{MD}_i)^2$ are asymptotically χ^2_p random variables, and $\mathrm{MED} = \sqrt{\chi^2_{p,0.5}}$ where $\chi^2_{p,0.5}$ is the median of the χ^2_p distribution. Since the target distribution is Gaussian, let

$$\mathrm{RD}_i = \frac{\sqrt{\chi^2_{p,0.5}}}{\mathrm{med}(D_i(A))} D_i(A) \quad \text{so that} \quad \tau = \frac{\sqrt{\chi^2_{\mathrm{p},0.5}}}{\mathrm{med}(\mathrm{D_i(A)})}. \tag{5.2}$$

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 $\mathrm{med}(\mathrm{MD}_i) \approx \mathrm{MED}$ where MED is the target population analog (obtained, for example, via simulation, or from the actual target distribution as in Equations (3.8), (3.9) and (3.10)), then use

$$\mathrm{RD}_i = \tau \; D_i(A) = \frac{\mathrm{MED}}{\mathrm{med}(D_i(A))} D_i(A). \tag{5.3}$$

The choice of the algorithm estimator $(T_A, \boldsymbol{C}_A)$ is important, and the $\sqrt{n}$ consistent RFCH estimator is a good choice. In this chapter we used the *R/Splus* function `cov.mcd` which is basically an implementation of the elemental FMCD 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 5.1. If $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ and an elemental FMCD 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 $n \to \infty$.

Table 5.1: **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

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

$$\left[\frac{\text{med}(D_i(A))}{\text{med}(\text{MD}_i)}\right]^2 \boldsymbol{C}_A \tag{5.4}$$

estimates $\text{Cov}(\boldsymbol{x})$. 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 5.1 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 the FCH starts to the 500 elemental starts. 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 (5.2). The results are shown in Table 5.1. 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 a hyperellipsoid, the DD plot can be used to examine which points are in the robust hyperellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - T_R)^T \boldsymbol{C}_R^{-1} (\boldsymbol{x} - T_R) \leq RD_{(h)}^2\} \tag{5.5}$$

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

$$\{\boldsymbol{x} : (\boldsymbol{x} - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x} - \overline{\boldsymbol{x}}) \leq MD_{(h)}^2\}. \tag{5.6}$$

In the DD plot, points below $RD_{(h)}$ correspond to cases that are in the hyperellipsoid given by Equation (5.5) while points to the left of $MD_{(h)}$ are in a hyperellipsoid determined by Equation (5.6).

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

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

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

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

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

Application 5.1. 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.

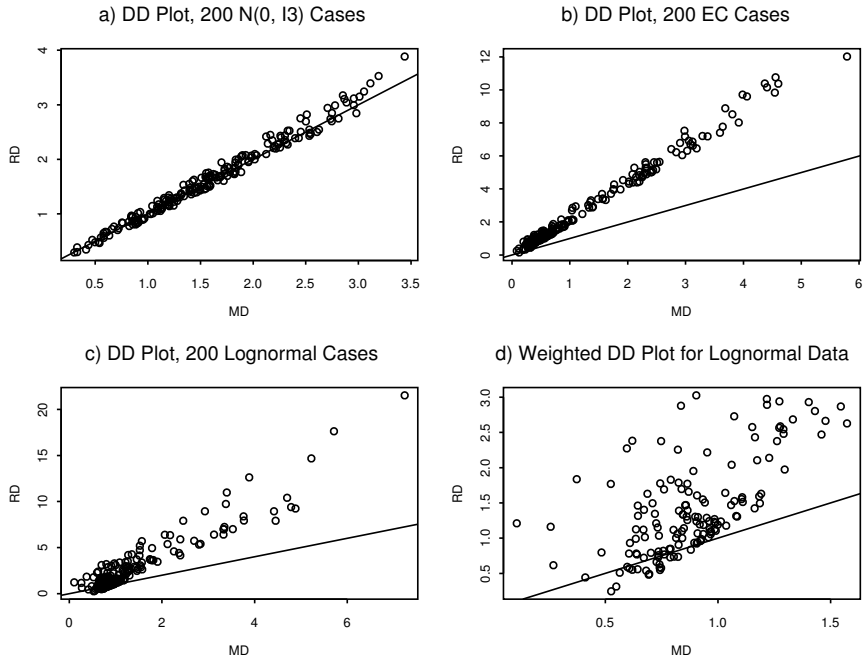

Figure 5.1: 4 DD Plots

Figure 5.1 shows the DD plots for 3 artificial data sets using `cov.mcd`. The DD plot for 200 $N_3(\mathbf{0}, \boldsymbol{I}_3)$ points shown in Figure 5.1a resembles the identity line. The DD plot for 200 points from the elliptically contoured distribution $0.6N_3(\mathbf{0}, \boldsymbol{I}_3) + 0.4N_3(\mathbf{0}, 25\ \boldsymbol{I}_3)$ in Figure 5.1b 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 $\mathrm{RD}_i \geq \sqrt{\chi^2_{p,.975}}$. 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 5.1 implies that the correlation of the 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 5.1b are highly correlated and still follow a line through the origin with a slope close to 2.0.

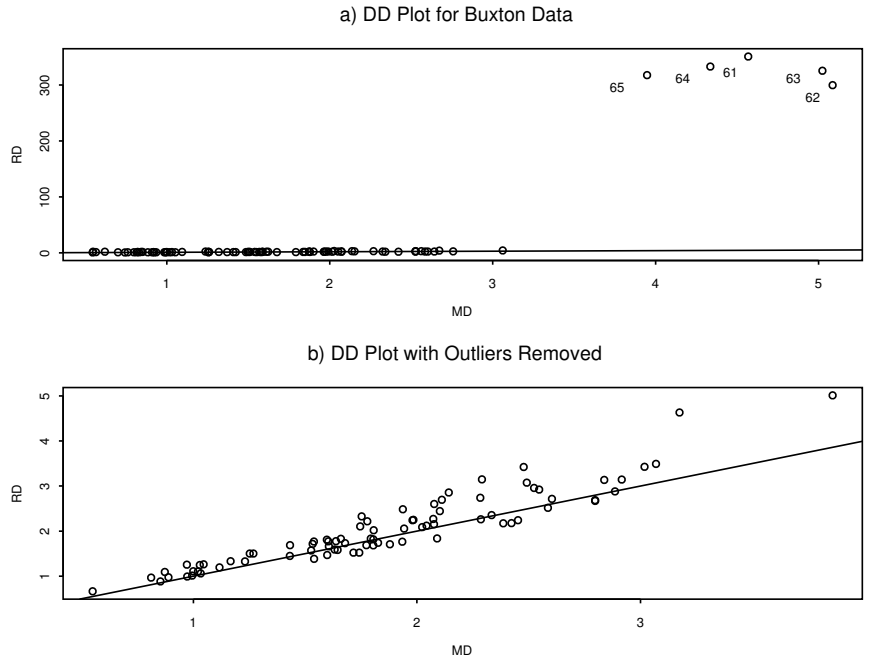

Figure 5.2: DD Plots for the Buxton Data

Figures 5.1c and 5.1d illustrate how to use the weighted DD plot. The ith case in Figure 5.1c is $(\exp(x_{i,1}), \exp(x_{i,2}), \exp(x_{i,3}))^T$ where $\boldsymbol{x}_i$ is the ith case in Figure 5.1a; 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 5.1d is the weighted DD plot where cases with $\mathrm{RD}_i \geq \sqrt{\chi^2_{3,.975}} \approx 3.06$ have been removed. Notice that the correlation of the plotted points is not close to one and that the best fitting line in Figure 5.1d may not pass through the origin. These results suggest that the distribution of $\boldsymbol{x}$ is not EC.

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 5.2. 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 5.2a shows the DD plot. Five head lengths were recorded to be around 5 feet and are massive outliers. Figure 5.2b 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 5.1a, then perhaps no transformation is needed. For the data in Figure 5.1c, a good numerical procedure should suggest coordinatewise log transforms. Following this transformation, the resulting plot shown in Figure 5.1a indicates that the transformation to normality was successful.

Application 5.2. The DD plot can be used to detect multivariate outliers. See Figures 4.2, 4.4, 5.2a and 5.7.

5.2 Robust Prediction Regions

Suppose that $(T_A, \boldsymbol{C}_A)$ is a good estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$. Section 5.1 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 (5.2). Then $(T_R, \boldsymbol{C}_R) \equiv (T_A, \boldsymbol{C}_A/\tau^2)$ is an estimator of multivariate location and dispersion.

Suppose $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}_M, b\ \boldsymbol{S}_M)$ is the sample mean and scaled sample covariance matrix applied to some subset of the data. The classical and RMVN estimators satisfy this assumption. For $h > 0$, the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}^2 \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\} \quad (5.7)$$

has volume equal to

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p \sqrt{det(\boldsymbol{C})} = \frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p b^{p/2} \sqrt{det(\boldsymbol{S}_M)}. \quad (5.8)$$

A future observation (random vector) $\boldsymbol{x}_f$ is in the region (5.7) if $D_{\boldsymbol{x}_f} \leq h$.

A large sample $(1-\alpha)100\%$ prediction region is a set $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n) \xrightarrow{P} 1 - \alpha$. Let $q_n = \min(1 - \alpha + 0.05, 1 - \alpha + p/n)$ for $\alpha > 0.1$ and

$$q_n = \min(1 - \alpha/2, 1 - \alpha + 10\alpha p/n), \quad \text{otherwise.}$$

$$\text{If } \ \mathrm{q_n} < 1 - \alpha + 0.001, \ \ \text{use } \ \mathrm{q_n} = 1 - \alpha. \quad (5.9)$$

If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$, then (5.7) is a large sample $(1 - \alpha)100\%$ prediction regions if $h = D_{(up)}$ where $D_{(up)}$ is the q_nth sample quantile of the D_i where the D_i^2 are given by (3.12). If $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ and $\boldsymbol{x}_f$ are iid from an EC distribution (with continuous decreasing g), then region (5.7) is asymptotically optimal in that its volume converges in probability to the volume of the minimum volume covering region $\{\boldsymbol{z} : (\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu}) \leq u_{1-\alpha}\}$ where $P(U \leq u_{1-\alpha}) = 1 - \alpha$ and U has pdf given by (3.10). The classical parametric MVN prediction region uses $MD_{\boldsymbol{x}_f} \leq \sqrt{\chi^2_{p,1-\alpha}}$.

Notice that for the data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, if $\boldsymbol{C}^{-1}$ exists, then $100q_n\%$ of the n cases are in the prediction region, and $q_n \to 1 - \alpha$ even if $(T, \boldsymbol{C})$ is not a good estimator. Hence the coverage q_n of the data is robust to model assumptions. Of course the volume of the prediction region could be large if a poor estimator $(T, \boldsymbol{C})$ is used or if the $\boldsymbol{x}_i$ do not come from an elliptically contoured distribution. Also notice that $q_n = 1 - \alpha/2$ or $q_n = 1 - \alpha + 0.05$

for $n \leq 20p$ and $q_n \to 1-\alpha$ as $n \to \infty$. If $q_n \equiv 1-\alpha$, then (5.7) is a large sample prediction region, but taking q_n given by (5.9) improves the finite sample performance of the region. Taking $q_n \equiv 1-\alpha$ does not take into account variability of $(T, \boldsymbol{C})$, and for small n the resulting prediction region tended to have undercoverage as high as $\min(0.05, \alpha/2)$. Using (5.9) helped reduce undercoverage for small n due to the unknown variability of $(T, \boldsymbol{C})$.

Three new prediction regions will be considered. The nonparametric region uses the classical estimator $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ and $h = D_{(up)}$. The semiparametric region uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ and $h = D_{(up)}$. The parametric MVN region uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ and $h^2 = \chi^2_{p,q_n}$ where $P(W \leq \chi^2_{p,\alpha}) = \alpha$ if $W \sim \chi^2_p$. All three regions are asymptotically optimal for MVN distributions with nonsingular $\boldsymbol{\Sigma}$. The first two regions are asymptotically optimal under the large class of EC distribution given by Assumption (E1) used in Theorem 4.8. For distributions with nonsingular covariance matrix $c_X\boldsymbol{\Sigma}$, the nonparametric region is a large sample $(1-\alpha)100\%$ prediction region, but regions with smaller volume may exist.

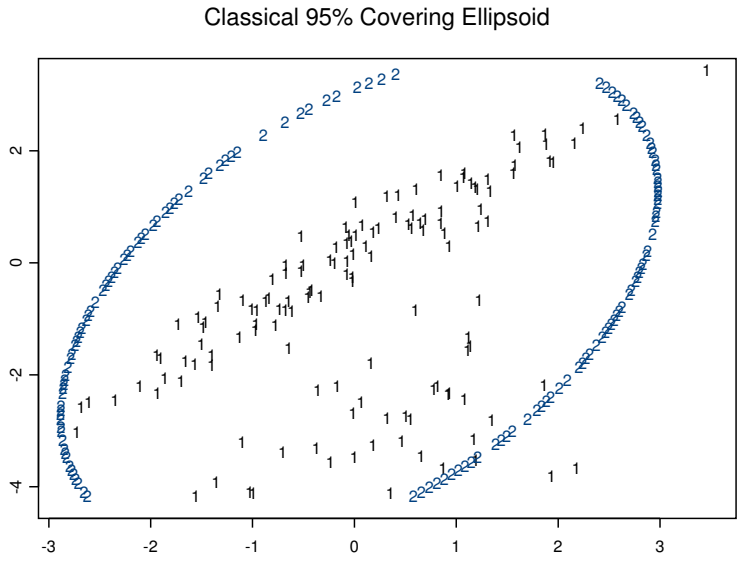

Figure 5.3: Artificial Bivariate Data

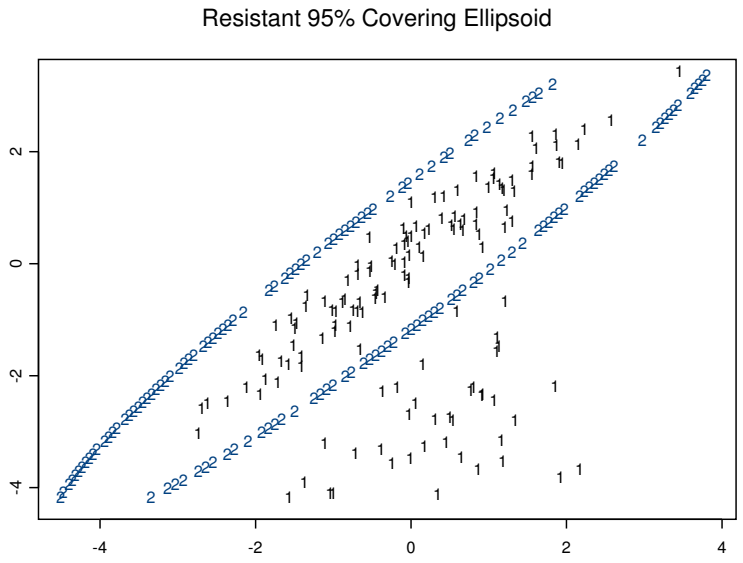

Figure 5.4: Artificial Data

Example 5.3. An artificial data set consisting of 100 iid cases from a

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

distribution and 40 iid cases from a bivariate normal distribution with mean $(0, -3)^T$ and covariance $\boldsymbol{I}_2$. Figure 5.3 shows the classical ellipsoid (with $MD \leq \sqrt{\chi^2_{2,0.95}}$) 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 parametric ellipsoid covers almost all of the data. Figure 5.4 displays the robust ellipsoid (using $RD \leq \sqrt{\chi^2_{2,0.95}}$) which contains most of the 100 "clean" cases and excludes the 40 outliers. Problem 5.5 recreates similar figures with the classical and RMVN estimators using $q_n = 0.95$.

Example 5.4. Buxton (1920) gives various measurements on 87 men including *height*, *head length*, *nasal height*, *bigonal breadth* and *cephalic index*. Five *heights* were recorded to be about 19mm and are massive outliers. First *height* and *nasal height* were used with $q_n = 0.95$. Figure 5.5 shows that the

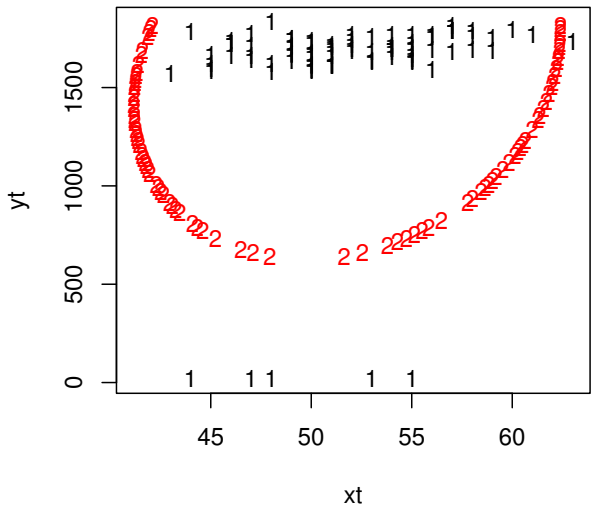

Figure 5.5: Ellipsoid is Inflated by Outliers

classical parametric prediction region (using $MD \leq \sqrt{\chi^2_{2,.95}}$) is quite large but does not include any of the outliers. Figure 5.6 shows that the parametric MVN prediction region (using $RD \leq \sqrt{\chi^2_{2,.95}}$) is not inflated by the outliers.

Next all 87 cases and 5 predictors were used. Figure 5.7 shows the RMVN DD plot with the identity line added as a visual aid. Points to the left of the vertical line are in the nonparametric large sample 90% prediction region. Points below the horizontal line are in the semiparametric region. The horizontal line at $RD = 3.33$ corresponding to the parametric MVN 90% region is obscured by the identity line. This region contains 78 of the cases. Since $n = 87$, the nonparametric and semiparametric regions used the 95th quantile. Since there were 5 outliers, this quantile was a linear combination of the largest clean distance and the smallest outlier distance. The semiparametric 90% region blows up unless the outlier proportion is small.

Figure 5.8 shows the DD plot and 3 prediction regions after the 5 outliers were removed. The classical and robust distances cluster about the identity line and the three regions are similar, with the parametric MVN region cutoff

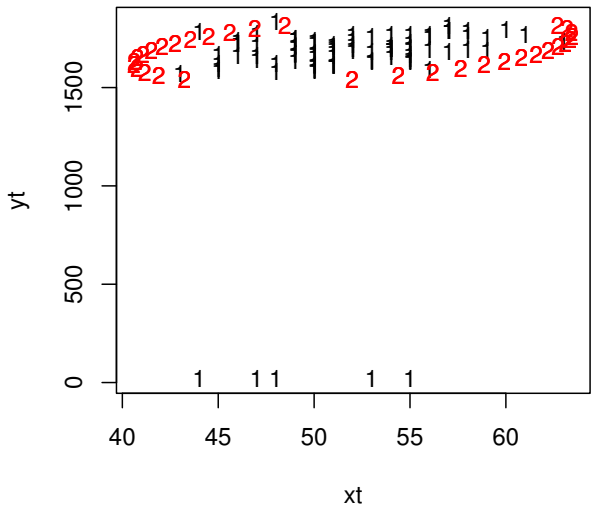

Figure 5.6: Ellipsoid Ignores Outliers

again at 3.33, slightly below the semiparametric region cutoff of 3.44.

Simulations for the prediction regions used $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{w}$ where $\boldsymbol{A} = diag(\sqrt{1}, ..., \sqrt{p})$, $\boldsymbol{w} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ (MVN), $\boldsymbol{w} \sim LN(\boldsymbol{0}, \boldsymbol{I}_p)$ where the marginals are iid lognormal(0,1), or $\boldsymbol{w} \sim MVT_p(1)$, a multivariate t distribution with 1 degree of freedom so the marginals are iid Cauchy(0,1). All simulations used 5000 runs and $\alpha = 0.1$.

For large n, the semiparametric and nonparametric regions are likely to have coverage near 0.90 because the coverage on the training sample is slightly larger than 0.9 and $\boldsymbol{x}_f$ comes from the same distribution as the $\boldsymbol{x}_i$. For $n = 10p$ and $2 \leq p \leq 40$, the semiparametric region had coverage near 0.9. The ratio of the volumes

$$\frac{h_i^p \sqrt{det(\boldsymbol{C}_i)}}{h_2^p \sqrt{det(\boldsymbol{C}_2)}}$$

was recorded where $i = 1$ was the nonparametric region, $i = 2$ was the semiparametric region, and $i = 3$ was the parametric MVN region. The volume ratio converges in probability to 1 for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ data, and the ratio converges to 1 for $i = 1$ if Assumption (E1) holds. The parametric MVN

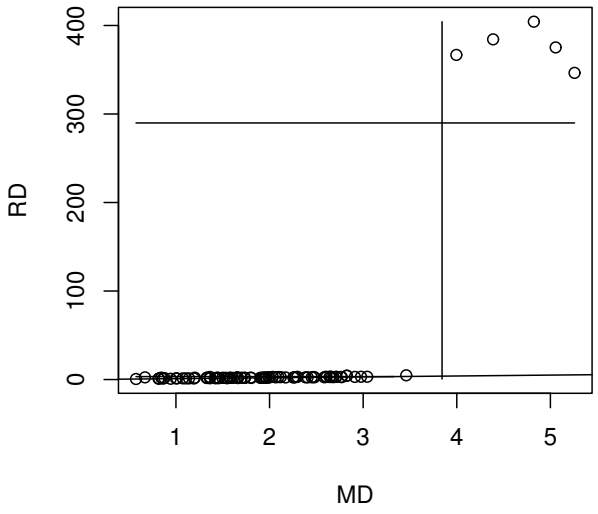

Figure 5.7: Prediction Regions for Buxton Data

region often had coverage much lower than 0.9 with a volume ratio near 0, recorded as 0+. The volume ratio tends to be tiny when the coverage is much less than the nominal value 0.9. For $10p \leq n \leq 20p$, the nonparametric region often had good coverage and volume ratio near 0.5.

Table 5.2: Coverages for 90% Prediction Regions

$\boldsymbol{w}$ dist	n	p	ncov	scov	mcov	voln	volm
MVN	600	30	0.906	0.919	0.902	0.503	0.512
MVN	1500	30	0.899	0.899	0.900	1.014	1.027
LN	1000	10	0.903	0.906	0.567	0.659	0+
MVT(1)	1000	10	0.914	0.914	0.541	22634.3	0+

Simulations and Table 5.2 suggest that for MVN data, the coverages (ncov, scov and mcov) for the 3 regions are near 90% for $n = 20p$ and that the volume ratios voln and volm are near 1 for $n = 50p$. With fewer than

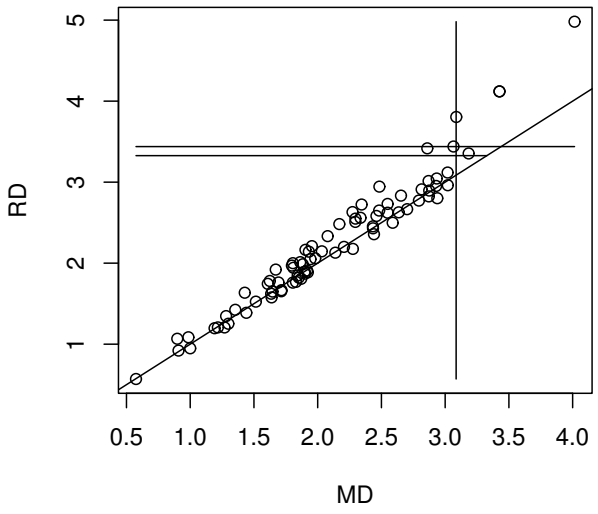

Figure 5.8: Prediction Regions for Buxton Data without Outliers

5000 runs, this result held for $2 \leq p \leq 80$. For the non–elliptically contoured LN data, the nonparametric region had voln well under 1, but the volume ratio blew up for $\boldsymbol{w} \sim MVT_p(1)$.

5.3 Summary

1) For $h > 0$, the hyperellipsoid $\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}$. A future observation (random vector) $\boldsymbol{x}_f$ is in this region if $D_{\boldsymbol{x}_f} \leq h$. A large sample $(1-\alpha)100\%$ prediction region is a set $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n) \xrightarrow{P} 1 - \alpha$ where $0 < \alpha < 1$.

2) The classical $(1-\alpha)100\%$ large sample prediction region is $\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq \chi^2_{p,1-\alpha}\}$ and works well if n is large and the data are iid MVN.

3) Let $q_n = \min(1 - \alpha + 0.05, 1 - \alpha + p/n)$ for $\alpha > 0.1$ and $q_n = \min(1 - \alpha/2, 1 - \alpha + 10\alpha p/n)$, otherwise. If $q_n < 1 - \alpha + 0.001$, set $q_n = 1 - \alpha$. If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$, then $\{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}$ is a large sample $(1-\alpha)100\%$ prediction regions if $h = D_{(up)}$ where $D_{(up)}$ is the q_nth sample quantile of the D_i. The nonparametric prediction region uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$

and the semiparametric prediction region uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$. The parametric MVN prediction region
$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(T, \boldsymbol{C}) \leq \chi^2_{p,q_n}\}$ also uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$.

4) These 3 regions can be displayed in an RMVN DD plot with cases in the nonparametric region corresponding to points to the left of the vertical line corresponding to $D_{(up)}(\overline{\boldsymbol{x}}, \boldsymbol{S})$. Cases in the semiparametric region correspond to points below the horizontal line corresponding to $D_{(up)}(T_{RMVN}, \boldsymbol{C}_{RMVN})$ while cases in the parametric MVN region correspond to points below the horizontal line corresponding to $\sqrt{\chi^2_{p,q_n}}$. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid with nonsingular covariance matrix $\boldsymbol{\Sigma}_x$. The three prediction regions are asymptotically optimal if the data is MVN. The semiparametric and nonparametric prediction regions are asymptotically optimal on a large class of EC distributions and the nonparametric prediction region is a large sample $100(1-\alpha)\%$ prediction region, although large sample prediction regions with smaller volume may exist.

5) Suppose m independent large sample $100(1-\alpha)\%$ prediction regions are made where $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid from the same distribution for each of the m runs. Let Y count the number of times $\boldsymbol{x}_f$ is in the prediction region. Then $Y \sim$ binomial $(m, 1-\alpha_n)$ where $1-\alpha_n$ is the true coverage and $1-\alpha_n \to 1-\alpha$ as $n \to \infty$. Simulation can be used to see if the true or actual coverage $1-\alpha_n$ is close to the nominal coverage $1-\alpha$. A prediction region with $1-\alpha_n < 1-\alpha$ is liberal and a region with $1-\alpha_n > 1-\alpha$ is conservative. It is better to be conservative by 5% than liberal by 5%. Parametric prediction regions tend to have large undercoverage and so are too liberal.

6) For prediction regions, want $n > 10p$ for the nonparametric prediction region and $n > 20p$ for the semiparametric prediction region.

5.4 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 normal-

ity while Koltchinskii and Li (1998) and Manzotti, Pérez and Quiroz (2002) have references for tests for elliptically contoured distributions.

There are few practical competitors for the Olive (2013b) prediction regions in Section 5.2. Parametric regions such as the classical region for multivariate normal data tend to have severe undercoverage because the data rarely follows the parametric distribution. Procedures that use brand name high breakdown multivariate location and dispersion estimators take too long to compute for $p > 2$.

5.5 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

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

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

R/Splus Problems

Warning: Use the command *source("G:/mpack.txt")* **to download the programs. See Preface or Section 15.2.** Typing the name of the `mpack` 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.

5.2. 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 RFCH DD plot of n $N_p(\mathbf{0}, \boldsymbol{I}_p)$ cases to 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 $p = 3$.)

5.3. 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 RFCH 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 $p = 3$.)

5.4***.** 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?

5.5. 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 RMVN covering ellipsoid. Include the two plots in *Word.*

```
> simx2 <- matrix(rnorm(200),nrow=100,ncol=2)
> outx2 <- matrix(10 + rnorm(80),nrow=40,ncol=2)
> outx2 <- rbind(outx2,simx2)
> ellipse(outx2)

> zout <- covrmvn(outx2)
> ellipse(outx2,center=zout$center,cov=zout$cov)
```

5.6. a) Download the function `mplot`.

b) Enter the commands in Problem 5.4b 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.

5.7 a) Download the function `wddplot`.

b) Enter the commands in Problem 5.4b to obtain a data set x.

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

5.8. Use the *R* command *source("G:/mrobdata.txt")* then *ddplot4(buxx,alpha=0.2)* and put the plot in *Word*. The Buxton data has 5 outliers, $p = 4$, and $n = 87$, so the 80% prediction regions use $1 - \alpha + p/n = 0.846$ percentiles. The output shows that the cutoffs are 2.527, 2.734 and 2.583 for the nonparametric, semiparametric and robust parametric prediction regions. The two horizontal lines that correspond to the robust distances are obscured by the identity line.

5.9. a) Type the *R* command `predsim()` and paste the output into *Word*.

This computes $\boldsymbol{x}_i \sim N_4(\boldsymbol{0}, diag(1, 2, 3, 4))$ for $i = 1, ..., 100$ and $\boldsymbol{x}_f = \boldsymbol{x}_{101}$. One hundred such data sets are made, and ncvr, scvr, mcvr counts the number of times $\boldsymbol{x}_f$ was in the nonparametric, semiparametric and parametric MVN 90% prediction regions. The volumes of the prediction regions are computed and voln, vols and volm are the average ratio of the volume of the ith prediction region over that of the semiparametric region. Hence vols is always equal to 1. For multivariate normal data, these ratios should converge to 1 as $n \to \infty$. Were the three coverages near 90%?

5.10. Tests for covariance matrices are very nonrobust to nonnormality. Let a plot of x versus y have x on the horizontal axis and y on the vertical axis. A good diagnostic is to use the DD plot. So a diagnostic for $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{\Sigma}_0$ is to plot $D_i(\overline{\boldsymbol{x}}, \boldsymbol{S})$ versus $D_i(\overline{\boldsymbol{x}}, \boldsymbol{\Sigma}_0)$ for $i = 1, ..., n$. If $n > 10p$ and H_0 is true, then the plotted points in the DD plot should cluster tightly about the identity line.

a) A test for sphericity is a test of $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}} = d\boldsymbol{I}_p$ for some unknown constant $d > 0$. Make a "DD plot" of $D_i^2(\overline{\boldsymbol{x}}, \boldsymbol{S})$ versus $D_i^2(\overline{\boldsymbol{x}}, \boldsymbol{I}_p)$. If $n > 10p$ and H_0 is true, then the plotted points in the "DD plot" should cluster tightly about the line through the origin with slope d. Use the *R* commands for this part and paste the plot into *Word*. The simulated data set has $\boldsymbol{x}_i \sim N_{10}(\boldsymbol{0}, 100\boldsymbol{I}_{10})$ where $n = 100$ and $p = 10$. Do the plotted points follow

a line through the origin with slope 100?

b) Now suppose there are k samples, and want to test $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}_1} = \cdots = \boldsymbol{\Sigma}_{\boldsymbol{x}_k}$, that is, all k populations have the same covariance matrix. As a diagnostic, make a DD plot of $D_i(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ versus $D_i(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_{pool})$ for $j = 1, ..., k$ and $i = 1, ..., n_i$. If each $n_i > 10p$ and H_0 is true, what line will the plotted points cluster about in each of the k DD plots?