p. 121 A <u>DD plot</u> is a plot of classical versus robust Mahalanobis distances $(MD_i$ vs $RD_i)$.

4.6 $MD_i = \sqrt{(\underline{x}_i - \bar{\underline{x}})^T S^{-1} (\underline{x}_i - \bar{\underline{x}})} = D_i(\bar{\underline{x}}, S)$.

4.7 Let (T_A, C_A) be a robust estimator,

$RD_i = \frac{\sqrt{\chi^2_{p,.5}}}{MED(D_i(A))} D_i(A)$ where $D_i(A) = D_i(T_A, C_A)$

and $\chi^2_{p,.5}$ is the median of the χ^2_p distribution.

Equivalently, $RD_i = D_i(T_R, C_R)$ where

$T_R = T_A$ and $C_R = \frac{C_A}{T^2}$ where

$T = \frac{\sqrt{\chi^2_{p,.5}}}{MED(D_i(A))}$ ← sample median of $D_1(A), \ldots, D_n(A)$

4.8 Suppose the data $\underline{x}_1, \ldots, \underline{x}_n$ are iid from a $EC_p(\underline{\mu}, \Sigma, g)$ distribution with 2nd moments. Then $(\bar{\underline{x}}, S)$ is a consistent estimator for $(\underline{\mu}, c_x \Sigma) = (\underline{\mu}, Cov(\underline{x}))$. Suppose that the robust estimator (T_A, C_A) is a consistent estimator for $(\underline{\mu}, a_A \Sigma)$ for some constant a_A.

Then $MD_i^2 \approx \frac{1}{c_x}(x_i-\mu)^T \Sigma^{-1}(x_i-\mu)$

$= \frac{a_A}{a_A} \frac{1}{c_x}(x_i-\mu)^T \Sigma^{-1}(x_i-\mu)$

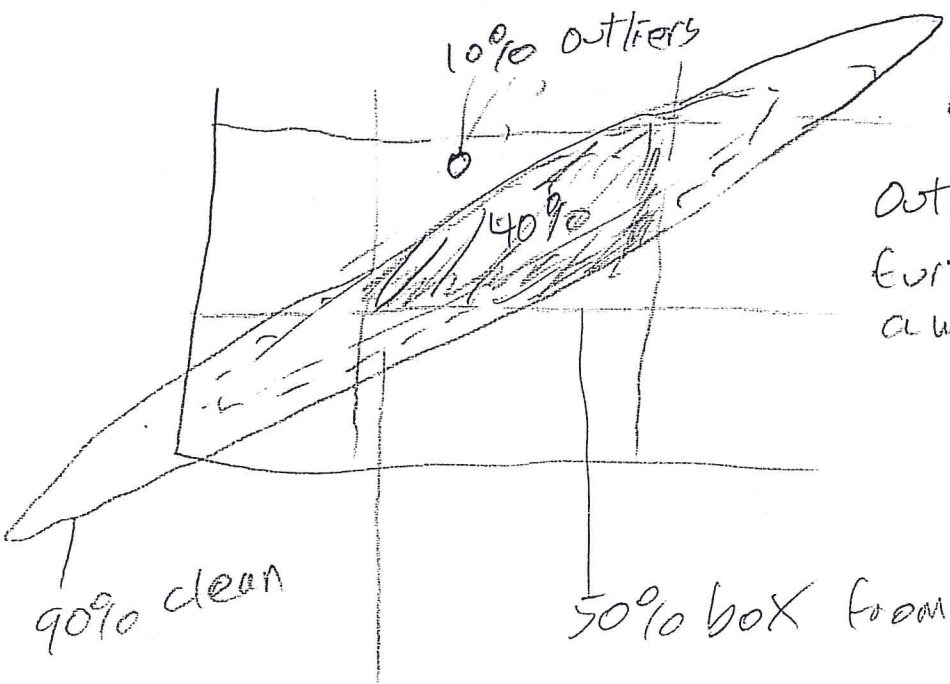
$= \frac{a_A}{c_x}(x_i-\mu)^T (a_A \Sigma)^{-1}(x_i-\mu) \approx \frac{a_A}{c_x} D_i^2(A)$.

Problem! $\frac{a_A}{c_x}$ depends on the EC distribution and

a_A, c_x and $\frac{a_A}{c_x}$ are usually unknown.

Strip ↓

(coordinatewise trimming: $p=1$ 25% trimmed mean uses a half set. As $p\uparrow$ need to decrease amount of trimming to get a half set. Also the method is poor for highly correlated data where $\Sigma \neq \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$:



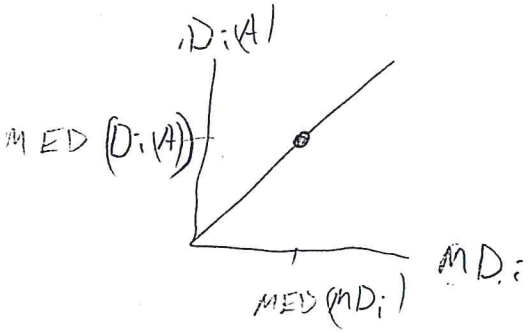
The 50% half set box has the outliers.

Outliers need to be further and further away as p increases.

50% hyperellipsoid does not contain the outliers.

↑ Strip

(49) p 123 A plot of MD_i vs $D_i(A)$ will have plotted points that cluster tightly about the line through $(0,0)^T$ and $(MED(MD_i), MED(D_i(A)))^T$. (The correlation goes to 1 as $n \to \infty$.)



Hence $D_i(A) \approx \frac{MED(D_i(A))}{MED(MD_i)} MD_i$.

(50) If $\underline{x}_1, \ldots, \underline{x}_n$ are iid $N_p(\mu, \Sigma)$, then

$\underbrace{MED(MD_i)}_{\text{sample median}} \approx \underbrace{\sqrt{\chi^2_{p,.5}}}_{\text{pop median}}$. Hence

$$D_i(A) \approx \frac{MED(D_i(A))}{\sqrt{\chi^2_{p,.5}}} MD_i$$

or $$MD_i \approx \frac{\sqrt{\chi^2_{p,.5}}}{MED(D_i(A))} D_i(A) = \gamma D_i(A) = RD_i.$$

(51) $RD_i = \gamma D_i(A) = \gamma D_i(T_A, C_A) =$

$$\gamma \sqrt{(x_i - T_R)^T C_A^{-1} (x_i - T_R)} = \quad \left(\gamma^2 C_A^{-1} = \left(\frac{C_A}{\gamma^2}\right)^{-1}\right)$$

$$\sqrt{(x_i - T_R)^T \left(\frac{C_A}{\gamma^2}\right)^{-1} (x_i - T_R)} = \sqrt{(x_i - T_R)^T C_R^{-1} (x_i - T_R)}$$

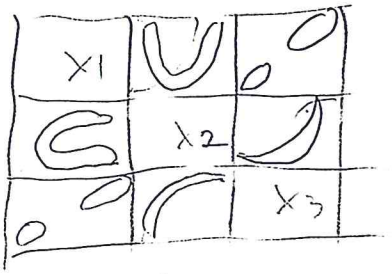
$$= D_i(T_R, C_R).$$

then the plotted points in the DD plot will cluster tightly about the identity line as $n \to \infty$. (25.5)

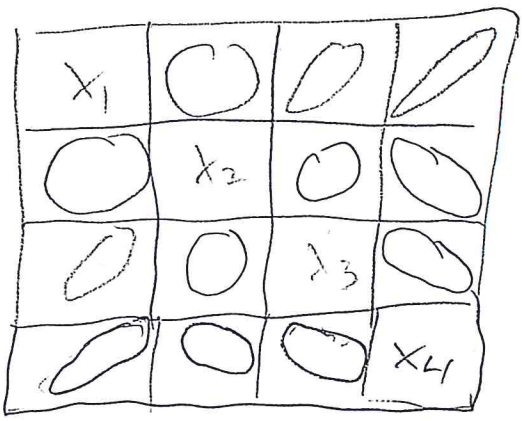
If the data are EC with nonsingular $\mathrm{Cov}(\underline{x})$, but not MVN, then the data will cluster tightly about some line through the origin, usually with slope > 1.

(5) Know If $\underline{x}_1, \ldots, \underline{x}_n$ are iid EC and $\underline{x} = (x_1, \ldots, x_p)^T$, then there should be no strong nonlinearities in a scatterplot matrix of $x_1, \ldots, x_p$.

(6)



not EC



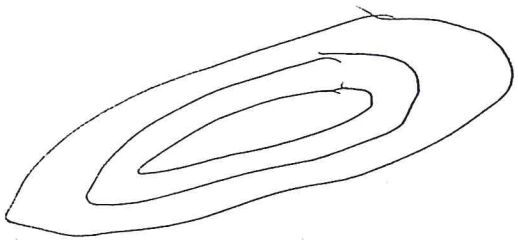
ex) MVN EC not MVN EC (See Fig 11.1.)

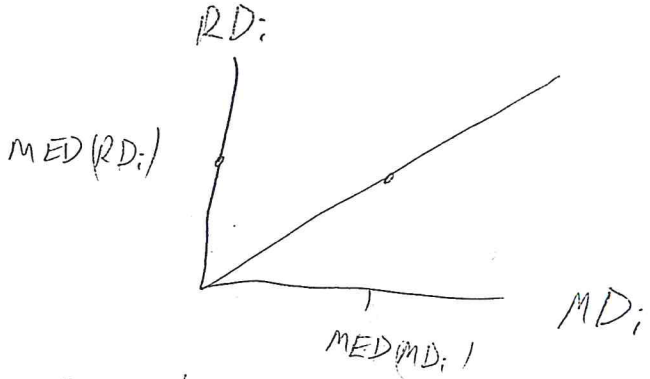
(7) For EC data, the ellipsoids

$$E_{RH} = \{\underline{z} \mid (\underline{z} - T_R)^T C_R^{-1} (\underline{z} - T_R) \le RD^2_{(h)}\} \text{ and}$$

$$E_{MH} = \{\underline{z} \mid (\underline{z} - \bar{\underline{x}})^T S^{-1} (\underline{z} - \bar{\underline{x}}) \le MD^2_{(h)}\}$$ approximate

the highest (eg 75% if $h \approx .95n$) density regions. Hence the 2 ellipsoids are approximately concentric.



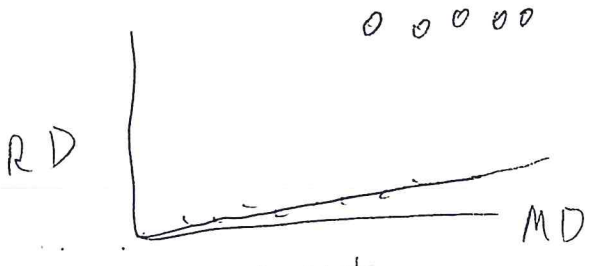


5.5) In the DD plot, points below $RD_{(h)}$ correspond to points in E_{RH} while points to the left of $MD_{(h)}$ correspond to points in E_{MH}.

5.6) know p 128 If (T_A, C_A) is a robust estimator, the DD plot is useful for detecting multivariate outliers.

ex) p 127-8 Buxton data x_1 = head length x_2 = nasal height x_3 = bigonal breadth x_4 = cephalic index

$\underline{x}_i = (x_{1i}, x_{2i}, x_{3i}, x_{4i})^T$, 5 head lengths were over 5 feet long.



DD plot

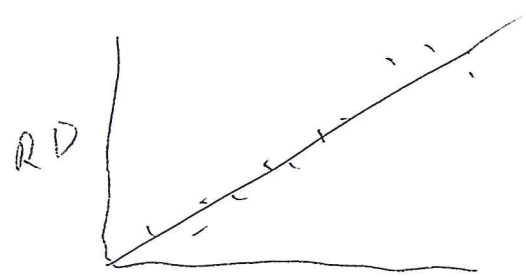
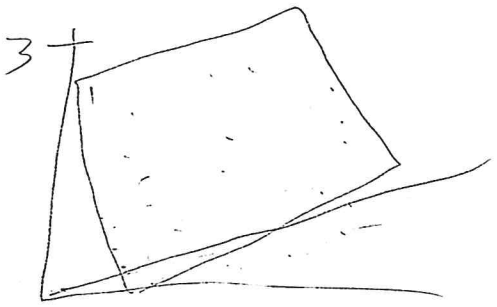


Fig 2 DD plot MD with outliers deleted

Southwest corner of the plot. This can be useful to see whether the data is tightly correlated about some line. 26.5

ex Fig 3.4 c) d) on P126

c)
looks fairly linear



weighted DD plot shows data is not tightly clustered about any line. The data dist is not EC.

§B.10 58) Outlier detection if $p > n$:

More than $\frac{n}{2}$ cases are in the bulk of the data:

a) Use Euclidean distances from the coordinatewise median $D_i(MED(W), I_p)$.

b) Let MED_j be the coordinatewise median computed from the cases x_i, with $D_i^2 \leq MED(D_i^2(MED_j, I_p))$ where $MED_0 = MED(W)$. often use $j = 0$ or 9.

Let $D_i = D_i(MED_j, I_p)$. Let 58327

$$w_i = \begin{cases} 1 & \text{if } D \le MED(D_1,\ldots,D_n) + 5\,MAD(D_1,\ldots,D_n) \\ 0 & \text{else} \end{cases}$$

The covmb2 set B consists of the $m \ge \frac{n}{2}$ cases with weight $w_i = 1$. The covmb2 estimator (T, C) has $T = \dfrac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$,

$$C = \frac{\sum_{i=1}^{n} w_i (x_i - T)(x_i - T)'}{\sum_{i=1}^{n} w_i - 1}$$

which is the sample mean and covariance matrix computed from the cases in B.

This is the concentration using hyperspheres: the dispersion matrix I_p does not change.

59] Can make a DD plot of $D_i(MED(w), I_p)$ vs $D_i(MED_j, I_p) = D_i(T_{COVMB2}, I_p)$

The plotted points tend to cluster about the identity line with outliers in the upper right corner of the plot with a gap between the bulk of the data and the outliers. To detect the outliers, the distance of the outliers from the bulk of the data increases roughly with $\sqrt{p}$.

60] RMVN! RMVN set U
RFCH! RFCH set V
covmb2! covmb2 set B

Let ...

contain $\geq \frac{n}{2}$ cases and (T_A, C_A)

$= (\bar{x}_A, S_A)$ is computed from those cases.

$(T_D, C_D) = (\bar{x}_A, d_A S_A)$ for $d_A > 0$, $d_A \neq 0$.

6) The two main methods to handle outliers are a) apply the classical statistical method to the cleaned data

b) plug in robust estimators for classical estimators: eg. (T, C) for $(\bar{x}, S)$.

62) The multivariate method applied to gets U and V is often the classical method applied to the cleaned cases in the scale and the plug in method using $(\bar{x}_A, S_A)$ instead of $(\bar{x}, S)$.

63) Resistant regression! Let $\underline{w}_i = (y_i, \underline{x}_i^T)^T$ and let the continuous predictors from $\underline{x}_i$ be $\underline{u}_i$. (the predictors that take on many values, so not gender). Apply the regression method to the m cases $\underline{w}_i$ corresponding to the set $D = \{i_1, \ldots, i_m\}$ applied to $\underline{u}_1, \ldots, \underline{u}_n$: MLR, GLMs, GAMs, LDA, QDA, k-NN etc.