Chapter 3
The Multivariate Location and Dispersion Model

This chapter describes the multivariate location and dispersion (MLD) model, random vectors, the population mean, the population covariance matrix, and the classical MLD estimators: the sample mean and the sample covariance matrix. Some important results on Mahalanobis distances and the volume of a hyperellipsoid are given. Robust MLD estimators are derived. The DD plot of classical versus robust Mahalanobis distances is used to detect outliers and to visualize practical prediction regions for a future test observation $\boldsymbol{x}_f$ that work even if the iid training data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ come from an unknown distribution.

The multivariate location and dispersion model is in many ways similar to the multiple linear regression model covered in Chapter 4. The data are iid vectors from some distribution such as the multivariate normal (MVN) distribution. The location parameter $\boldsymbol{\mu}$ of interest may be the mean or the center of symmetry of an elliptically contoured distribution. Hyperellipsoids will be estimated instead of hyperplanes, and Mahalanobis distances will be used instead of absolute residuals to determine if an observation is a potential outlier.

Definition 3.1. An important *multivariate location and dispersion model* is $\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{e}$ where $\boldsymbol{Y}$ and $\boldsymbol{e}$ are $p \times 1$ random vectors, while $\boldsymbol{\mu}$ is a $p \times 1$ population *location* vector. Often the $\boldsymbol{e}_i$ are iid with a $p \times p$ symmetric positive definite population *dispersion* matrix $\boldsymbol{\Sigma}$. An important parametric multivariate location and dispersion model is a joint distribution with joint pdf $f(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for a $p \times 1$ random vector $\boldsymbol{x}$ where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are as above. Thus $P(\boldsymbol{x} \in A) = \int_A f(\boldsymbol{z}) d\boldsymbol{z}$ for suitable sets A.

Notation: Usually a vector $\boldsymbol{x}$ will be column vector, and a row vector $\boldsymbol{x}^T$ will be the transpose of the vector $\boldsymbol{x}$. However,

$$\int_A f(\boldsymbol{z}) d\boldsymbol{z} = \int_A f(z_1, ..., z_p) dz_1 \cdots dz_p.$$

The notation $f(z_1, ..., z_p)$ will be used to write out the components z_i of a joint pdf $f(\boldsymbol{z})$ although in the formula for the pdf, e.g. $f(\boldsymbol{z}) = c\exp(\boldsymbol{z}^T\boldsymbol{z})$, $\boldsymbol{z}$ is a column vector.

Definition 3.2. A $p \times 1$ *random vector* $\boldsymbol{x} = (x_1, ..., x_p)^T = (X_1, ..., X_p)^T$ where $X_1, ..., X_p$ are p random variables. A *case* or *observation* consists of the p random variables measured for one person or thing. For multivariate location and dispersion the ith case is $\boldsymbol{x}_i = (x_{i,1}, ..., x_{i,p})^T$. There are n cases, and context will be used to determine whether $\boldsymbol{x}$ is the random vector or the observed value of the random vector. *Outliers* are cases that lie far away from the bulk of the data, and they can ruin a classical analysis.

Assume that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are n iid $p \times 1$ random vectors and that the joint pdf of $\boldsymbol{x}_i$ is $f(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Also assume that the data $\boldsymbol{x}_i$ has been observed and stored in an $n \times p$ matrix

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_p \end{bmatrix}$$

where the ith row of $\boldsymbol{W}$ is the ith case $\boldsymbol{x}_i^T$ and the jth column $\boldsymbol{v}_j$ of $\boldsymbol{W}$ corresponds to n measurements of the jth random variable X_j for $j = 1, ..., p$. Hence the n rows of the data matrix $\boldsymbol{W}$ correspond to the n cases, while the p columns correspond to measurements on the p random variables $X_1, ..., X_p$. For example, the data may consist of n visitors to a hospital where the $p = 2$ variables *height* and *weight* of each individual were measured.

Notation: *In the theoretical sections of this text, $\boldsymbol{x}_i$ will sometimes be a random vector and sometimes the observed data.* Some texts, for example Johnson and Wichern (1988, pp. 7, 53), use $\boldsymbol{X}$ to denote the $n \times p$ data matrix and an $n \times 1$ random vector, relying on the context to indicate whether $\boldsymbol{X}$ is a random vector or data matrix. Software tends to use different notation. For example, R will use commands such as

```
var(x)
```

to compute the sample covariance matrix of the data. Hence x corresponds to $\boldsymbol{W}$, x[,1] is the first column of x, and $x[4,]$ is the 4th row of x.

The next two sections consider elliptically contoured distributions, including the multivariate normal distribution. These distributions are important models for multivariate data. Although usually random vectors in this text are denoted by $\boldsymbol{x}$, $\boldsymbol{y}$, or $\boldsymbol{z}$, the next two sections will usually use the notation $\boldsymbol{X} = (X_1, ..., X_p)^T$ and $\boldsymbol{Y}$ for the random vectors, and $\boldsymbol{x} = (x_1, ..., x_p)^T$ for the observed value of the random vector. This notation will be useful to avoid confusion when studying conditional distributions such as $\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}$.

3.1 The Multivariate Normal Distribution

Definition 3.3: Rao (1965, p. 437). A $p \times 1$ random vector $\boldsymbol{X}$ has a $p-$dimensional *multivariate normal distribution* $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ iff $\boldsymbol{t}^T\boldsymbol{X}$ has a univariate normal distribution for any $p \times 1$ vector $\boldsymbol{t}$.

If $\boldsymbol{\Sigma}$ is positive definite, then $\boldsymbol{X}$ has a pdf

$$f(\boldsymbol{z}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(1/2)(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})} \tag{3.1}$$

where $|\boldsymbol{\Sigma}|^{1/2}$ is the square root of the determinant of $\boldsymbol{\Sigma}$. Note that if $p = 1$, then the quadratic form in the exponent is $(z-\mu)(\sigma^2)^{-1}(z-\mu)$ and X has the univariate $N(\mu, \sigma^2)$ pdf. If $\boldsymbol{\Sigma}$ is positive semidefinite but not positive definite, then $\boldsymbol{X}$ has a degenerate distribution. For example, the univariate $N(0, 0^2)$ distribution is degenerate (the point mass at 0).

Definition 3.4. If second moments exist, the *population mean* of a random $p \times 1$ vector $\boldsymbol{X} = (X_1, ..., X_p)^T$ is

$$E(\boldsymbol{X}) = (E(X_1), ..., E(X_p))^T$$

and the $p \times p$ *population covariance matrix*

$$\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{X}} = E(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{X} - E(\boldsymbol{X}))^T = (\sigma_{ij}) = (\sigma_{i,j}).$$

That is, the ij entry of $\text{Cov}(\boldsymbol{X})$ is $\text{Cov}(X_i, X_j) = \sigma_{i,j} = \sigma_{ij}$.

The covariance matrix is also called the variance–covariance matrix and variance matrix. Sometimes the notation $\text{Var}(\boldsymbol{X})$ is used. Note that $\text{Cov}(\boldsymbol{X})$ is a symmetric positive semidefinite matrix. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $p \times 1$ random vectors, $\boldsymbol{a}$ a conformable constant vector and $\boldsymbol{A}$ and $\boldsymbol{B}$ are conformable constant matrices, then

$$E(\boldsymbol{a} + \boldsymbol{X}) = \boldsymbol{a} + E(\boldsymbol{X}) \quad \text{and} \quad E(\boldsymbol{X} + \boldsymbol{Y}) = E(\boldsymbol{X}) + E(\boldsymbol{Y}) \tag{3.2}$$

and

$$E(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}E(\boldsymbol{X}) \quad \text{and} \quad E(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}) = \boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}. \tag{3.3}$$

Thus

$$\text{Cov}(\boldsymbol{a} + \boldsymbol{A}\boldsymbol{X}) = \text{Cov}(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}\text{Cov}(\boldsymbol{X})\boldsymbol{A}^T. \tag{3.4}$$

Some important properties of multivariate normal (MVN) distributions are given in the following three theorems. These theorems can be proved using results from Johnson and Wichern (1988, p. 127-132) or Severini (2005, ch. 8).

Theorem 3.1. a) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and

$$\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{X}} = \boldsymbol{\Sigma}.$$

b) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then any linear combination $\boldsymbol{t}^T\boldsymbol{X} = t_1X_1 + \cdots + t_pX_p \sim N_1(\boldsymbol{t}^T\boldsymbol{\mu}, \boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$. Conversely, if $\boldsymbol{t}^T\boldsymbol{X} \sim N_1(\boldsymbol{t}^T\boldsymbol{\mu}, \boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$ for every $p \times 1$ vector $\boldsymbol{t}$, then $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

c) **The joint distribution of independent normal random variables is MVN.** If $X_1, ..., X_p$ are independent univariate normal $N(\mu_i, \sigma_i^2)$ random vectors, then $\boldsymbol{X} = (X_1, ..., X_p)^T$ is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = (\mu_1, ..., \mu_p)$ and $\boldsymbol{\Sigma} = diag(\sigma_1^2, ..., \sigma_p^2)$ (so the off diagonal entries $\sigma_{ij} = 0$ while the diagonal entries of $\boldsymbol{\Sigma}$ are $\sigma_{ii} = \sigma_i^2$).

d) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and if $\boldsymbol{A}$ is a $q \times p$ matrix, then $\boldsymbol{A}\boldsymbol{X} \sim N_q(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. If $\boldsymbol{a}$ is a $p \times 1$ vector of constants and b is a constant, then $\boldsymbol{a} + b\boldsymbol{X} \sim N_p(\boldsymbol{a} + b\boldsymbol{\mu}, b^2\boldsymbol{\Sigma})$. (Note that $b\boldsymbol{X} = b\boldsymbol{I}_p\boldsymbol{X}$ with $\boldsymbol{A} = b\boldsymbol{I}_p$.)

It will be useful to partition $\boldsymbol{X}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$. Let $\boldsymbol{X}_1$ and $\boldsymbol{\mu}_1$ be $q \times 1$ vectors, let $\boldsymbol{X}_2$ and $\boldsymbol{\mu}_2$ be $(p-q) \times 1$ vectors, let $\boldsymbol{\Sigma}_{11}$ be a $q \times q$ matrix, let $\boldsymbol{\Sigma}_{12}$ be a $q \times (p-q)$ matrix, let $\boldsymbol{\Sigma}_{21}$ be a $(p-q) \times q$ matrix, and let $\boldsymbol{\Sigma}_{22}$ be a $(p-q) \times (p-q)$ matrix. Then

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Theorem 3.2. a) **All subsets of a MVN are MVN:** $(X_{k_1}, ..., X_{k_q})^T \sim N_q(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ where $\tilde{\boldsymbol{\mu}}_i = E(X_{k_i})$ and $\tilde{\boldsymbol{\Sigma}}_{ij} = \text{Cov}(X_{k_i}, X_{k_j})$. In particular, $\boldsymbol{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\boldsymbol{X}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$.

b) If $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent, then $\text{Cov}(\boldsymbol{X}_1, \boldsymbol{X}_2) = \boldsymbol{\Sigma}_{12} = E[(\boldsymbol{X}_1 - E(\boldsymbol{X}_1))(\boldsymbol{X}_2 - E(\boldsymbol{X}_2))^T] = \boldsymbol{0}$, a $q \times (p-q)$ matrix of zeroes.

c) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent iff $\boldsymbol{\Sigma}_{12} = \boldsymbol{0}$.

d) If $\boldsymbol{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\boldsymbol{X}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$ are independent, then

$$\begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} \sim N_p\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right).$$

Theorem 3.3. The conditional distribution of a MVN is MVN. If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal with mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. That is,

$$\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2 \sim N_q(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

Example 3.1. Let $p = 2$ and let $(Y, X)^T$ have a bivariate normal distribution. That is,

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \text{Cov}(Y,X) \\ \text{Cov}(X,Y) & \sigma_X^2 \end{pmatrix} \right).$$

Also the population correlation between X and Y is given by

$$\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{VAR}(X)}\sqrt{\text{VAR}(Y)}} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

if $\sigma_X > 0$ and $\sigma_Y > 0$. Then $Y|X = x \sim N(E(Y|X = x), \text{VAR}(Y|X = x))$ where the conditional mean

$$E(Y|X = x) = \mu_Y + \text{Cov}(Y,X)\frac{1}{\sigma_X^2}(x - \mu_X) = \mu_Y + \rho(X,Y)\sqrt{\frac{\sigma_Y^2}{\sigma_X^2}}(x - \mu_X)$$

and the conditional variance

$$\text{VAR}(Y|X = x) = \sigma_Y^2 - \text{Cov}(X,Y)\frac{1}{\sigma_X^2}\text{Cov}(X,Y) =$$

$$\sigma_Y^2 - \rho(X,Y)\sqrt{\frac{\sigma_Y^2}{\sigma_X^2}}\rho(X,Y)\sqrt{\sigma_X^2}\sqrt{\sigma_Y^2} = \sigma_Y^2 - \rho^2(X,Y)\sigma_Y^2 = \sigma_Y^2[1 - \rho^2(X,Y)].$$

Also $aX + bY$ is univariate normal with mean $a\mu_X + b\mu_Y$ and variance

$$a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\ \text{Cov}(X,Y).$$

Remark 3.1. There are several common misconceptions. First, **it is not true that every linear combination $\boldsymbol{t}^T\boldsymbol{X}$ of normal random variables is a normal random variable,** and **it is not true that all uncorrelated normal random variables are independent**. The key condition in Theorem 3.1b and Theorem 3.2c is that the joint distribution of $\boldsymbol{X}$ is MVN. It is possible that $X_1, X_2, ..., X_p$ each has a marginal distribution that is univariate normal, but the joint distribution of $\boldsymbol{X}$ is not MVN. Examine the following example from Rohatgi (1976, p. 229). Suppose that the joint pdf of X and Y is a mixture of two bivariate normal distributions both with $EX = EY = 0$ and $\text{VAR}(X) = \text{VAR}(Y) = 1$, but $\text{Cov}(X,Y) = \pm\rho$. Hence $f(x,y) =$

$$\frac{1}{2}\ \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left(\frac{-1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right) +$$

$$\frac{1}{2}\ \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left(\frac{-1}{2(1-\rho^2)}(x^2 + 2\rho xy + y^2)\right) \equiv \frac{1}{2}f_1(x,y) + \frac{1}{2}f_2(x,y)$$

where x and y are real and $0 < \rho < 1$. Since both marginal distributions of $f_i(x,y)$ are N(0,1) for $i = 1$ and 2 by Theorem 3.2 a), the marginal distributions of X and Y are N(0,1). Since $\int\int xyf_i(x,y)dxdy = \rho$ for $i = 1$ and $-\rho$

for $i = 2$, X and Y are uncorrelated, but X and Y are not independent since $f(x, y) \neq f_X(x) f_Y(y)$.

Remark 3.2. In Theorem 3.3, suppose that $\boldsymbol{X} = (Y, X_2, ..., X_p)^T$. Let $X_1 = Y$ and $\boldsymbol{X}_2 = (X_2, ..., X_p)^T$. Then $E[Y|\boldsymbol{X}_2] = \beta_1 + \beta_2 X_2 + \cdots + \beta_p X_p$ and $\mathrm{VAR}[Y|\boldsymbol{X}_2]$ is a constant that does not depend on $\boldsymbol{X}_2$. Hence $Y = \beta_1 + \beta_2 X_2 + \cdots + \beta_p X_p + e$ follows the multiple linear regression model.

3.2 Elliptically Contoured Distributions

Definition 3.5: Johnson (1987, p. 107-108). A $p \times 1$ random vector $\boldsymbol{X}$ has an *elliptically contoured distribution,* also called an *elliptically symmetric distribution,* if $\boldsymbol{X}$ has joint pdf

$$f(\boldsymbol{z}) = k_p |\boldsymbol{\Sigma}|^{-1/2} g[(\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu})], \tag{3.5}$$

and we say $\boldsymbol{X}$ has an elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution.

If $\boldsymbol{X}$ has an elliptically contoured (EC) distribution, then the characteristic function of $\boldsymbol{X}$ is

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp(i \boldsymbol{t}^T \boldsymbol{\mu}) \psi(\boldsymbol{t}^T \boldsymbol{\Sigma} \boldsymbol{t}) \tag{3.6}$$

for some function ψ. If the second moments exist, then

$$E(\boldsymbol{X}) = \boldsymbol{\mu} \tag{3.7}$$

and

$$\mathrm{Cov}(\boldsymbol{X}) = c_X \boldsymbol{\Sigma} \tag{3.8}$$

where $c_X = -2\psi'(0)$.

Definition 3.6. The *population squared Mahalanobis distance*

$$U \equiv D^2 = D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}). \tag{3.9}$$

For elliptically contoured distributions, U has pdf

$$h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)} k_p u^{p/2 - 1} g(u). \tag{3.10}$$

For $c > 0$, an $EC_p(\boldsymbol{\mu}, c\boldsymbol{I}, g)$ distribution is *spherical about* $\boldsymbol{\mu}$ where $\boldsymbol{I}$ is the $p \times p$ identity matrix. The *multivariate normal distribution* $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has $k_p = (2\pi)^{-p/2}$, $\psi(u) = g(u) = \exp(-u/2)$ and $h(u)$ is the χ^2_p pdf. The following theorem is useful for proving properties of EC distributions without using the characteristic function (3.6). See Eaton (1986) and Cook (1998a, p. 57, 130).

Theorem 3.4. Let $\boldsymbol{X}$ be a $p \times 1$ random vector with 1st moments; i.e., $E(\boldsymbol{X})$ exists. Let $\boldsymbol{B}$ be any constant full rank $p \times r$ matrix where $1 \leq r \leq p$. Then $\boldsymbol{X}$ is elliptically contoured iff for all such conforming matrices $\boldsymbol{B}$,

$$E(\boldsymbol{X}|\boldsymbol{B}^T\ \boldsymbol{X}) = \boldsymbol{\mu} + \boldsymbol{M}_B\boldsymbol{B}^T(\boldsymbol{X} - \boldsymbol{\mu}) = \boldsymbol{a}_B + \boldsymbol{M}_B\boldsymbol{B}^T\boldsymbol{X} \tag{3.11}$$

where the $p \times 1$ constant vector $\boldsymbol{a}_B$ and the $p \times r$ constant matrix $\boldsymbol{M}_B$ both depend on $\boldsymbol{B}$.

A useful fact is that $\boldsymbol{a}_B$ and $\boldsymbol{M}_B$ do not depend on g:

$$\boldsymbol{a}_B = \boldsymbol{\mu} - \boldsymbol{M}_B\boldsymbol{B}^T\boldsymbol{\mu} = (\boldsymbol{I}_p - \boldsymbol{M}_B\boldsymbol{B}^T)\boldsymbol{\mu},$$

and

$$\boldsymbol{M}_B = \boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B})^{-1}.$$

See Problem 3.11. Notice that in the formula for $\boldsymbol{M}_B$, $\boldsymbol{\Sigma}$ can be replaced by $c\boldsymbol{\Sigma}$ where $c > 0$ is a constant. In particular, if the EC distribution has second moments, $\mathrm{Cov}(\boldsymbol{X})$ can be used instead of $\boldsymbol{\Sigma}$.

To use Theorem 3.4 to prove interesting properties, partition $\boldsymbol{X}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ as above Theorem 3.2. Also assume that the $(p+1) \times 1$ vector $(Y, \boldsymbol{X}^T)^T$ is $EC_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ where Y is a random variable, $\boldsymbol{X}$ is a $p \times 1$ vector, and use

$$\begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix},\ \boldsymbol{\mu} = \begin{pmatrix} \mu_Y \\ \boldsymbol{\mu}_X \end{pmatrix},\ \text{and}\ \boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{YY} & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XX} \end{pmatrix}.$$

Theorem 3.5. Let $\boldsymbol{X} \sim EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ and assume that $E(\boldsymbol{X})$ exists.

a) Any subset of $\boldsymbol{X}$ is EC, in particular $\boldsymbol{X}_1$ is EC.

b) (Cook 1998a p. 131, Kelker 1970). If $\mathrm{Cov}(\boldsymbol{X})$ is nonsingular,

$$\mathrm{Cov}(\boldsymbol{X}|\boldsymbol{B}^T\boldsymbol{X}) = d_g(\boldsymbol{B}^T\boldsymbol{X})[\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{\Sigma}]$$

where the real valued function $d_g(\boldsymbol{B}^T\boldsymbol{X})$ is constant iff $\boldsymbol{X}$ is MVN.

Proof of a). Let $\boldsymbol{A}$ be an arbitrary full rank $q \times r$ matrix where $1 \leq r \leq q$. Let

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{0} \end{pmatrix}.$$

Then $\boldsymbol{B}^T\boldsymbol{X} = \boldsymbol{A}^T\boldsymbol{X}_1$, and

$$E[\boldsymbol{X}|\boldsymbol{B}^T\boldsymbol{X}] = E\left[\begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}|\boldsymbol{A}^T\boldsymbol{X}_1\right] =$$

$$\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{M}_{1B} \\ \boldsymbol{M}_{2B} \end{pmatrix} (\boldsymbol{A}^T\ \boldsymbol{0}^T) \begin{pmatrix} \boldsymbol{X}_1 - \boldsymbol{\mu}_1 \\ \boldsymbol{X}_2 - \boldsymbol{\mu}_2 \end{pmatrix}$$

by Theorem 3.4. Hence $E[\boldsymbol{X}_1|\boldsymbol{A}^T\boldsymbol{X}_1] = \boldsymbol{\mu}_1 + \boldsymbol{M}_{1B}\boldsymbol{A}^T(\boldsymbol{X}_1 - \boldsymbol{\mu}_1)$. Since $\boldsymbol{A}$ was arbitrary, $\boldsymbol{X}_1$ is EC by Theorem 3.4. Notice that $\boldsymbol{M}_B = \boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B})^{-1} =$

$$\begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{0} \end{pmatrix} \left[(\boldsymbol{A}^T \ \boldsymbol{0}^T) \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{0} \end{pmatrix} \right]^{-1} = \begin{pmatrix} \boldsymbol{M}_{1B} \\ \boldsymbol{M}_{2B} \end{pmatrix}.$$

Hence

$$\boldsymbol{M}_{1B} = \boldsymbol{\Sigma}_{11}\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{\Sigma}_{11}\boldsymbol{A})^{-1}$$

and $\boldsymbol{X}_1$ is EC with location and dispersion parameters $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_{11}$. □

Theorem 3.6. Let $(Y, \boldsymbol{X}^T)^T$ be $EC_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ where Y is a random variable.

a) Assume that $E[(Y, \boldsymbol{X}^T)^T]$ exists. Then $E(Y|\boldsymbol{X}) = \alpha + \boldsymbol{\beta}^T\boldsymbol{X}$ where $\alpha = \mu_Y - \boldsymbol{\beta}^T\boldsymbol{\mu}_X$ and

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}.$$

b) Even if the first moment does not exist, the conditional median

$$\text{MED}(Y|\boldsymbol{X}) = \alpha + \boldsymbol{\beta}^T\boldsymbol{X}$$

where α and $\boldsymbol{\beta}$ are given in a).

Proof. a) The trick is to choose $\boldsymbol{B}$ so that Theorem 3.4 applies. Let

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{0}^T \\ \boldsymbol{I}_p \end{pmatrix}.$$

Then $\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B} = \boldsymbol{\Sigma}_{XX}$ and

$$\boldsymbol{\Sigma}\boldsymbol{B} = \begin{pmatrix} \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XX} \end{pmatrix}.$$

$$\text{Now} \ \ \text{E}\left[\begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix} \mid \boldsymbol{X} \right] = \text{E}\left[\begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix} \mid \boldsymbol{B}^{\text{T}} \begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix} \right]$$

$$= \boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B})^{-1}\boldsymbol{B}^T \begin{pmatrix} Y - \mu_Y \\ \boldsymbol{X} - \boldsymbol{\mu}_X \end{pmatrix}$$

by Theorem 3.4. The right hand side of the last equation is equal to

$$\boldsymbol{\mu} + \begin{pmatrix} \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XX} \end{pmatrix} \boldsymbol{\Sigma}_{XX}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_X) = \begin{pmatrix} \mu_Y - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\mu}_X + \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{X} \\ \boldsymbol{X} \end{pmatrix}$$

and the result follows since $\boldsymbol{\beta}^T = \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}$.

b) See Croux et al. (2001) for references.

Example 3.2. This example illustrates another application of Theorem 3.4. Suppose that $\boldsymbol{X}$ comes from a mixture of two multivariate normals with

the same mean and proportional covariance matrices. That is, let

$$\boldsymbol{X} \sim (1-\gamma)N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \gamma N_p(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$$

where $c > 0$ and $0 < \gamma < 1$. Since the multivariate normal distribution is elliptically contoured (and see Theorem 11.1c),

$$E(\boldsymbol{X}|\boldsymbol{B}^T\boldsymbol{X}) = (1-\gamma)[\boldsymbol{\mu} + \boldsymbol{M}_1\boldsymbol{B}^T(\boldsymbol{X}-\boldsymbol{\mu})] + \gamma[\boldsymbol{\mu} + \boldsymbol{M}_2\boldsymbol{B}^T(\boldsymbol{X}-\boldsymbol{\mu})]$$

$$= \boldsymbol{\mu} + [(1-\gamma)\boldsymbol{M}_1 + \gamma\boldsymbol{M}_2]\boldsymbol{B}^T(\boldsymbol{X}-\boldsymbol{\mu}) \equiv \boldsymbol{\mu} + \boldsymbol{M}\boldsymbol{B}^T(\boldsymbol{X}-\boldsymbol{\mu}).$$

Since $\boldsymbol{M}_B$ only depends on $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$, it follows that $\boldsymbol{M}_1 = \boldsymbol{M}_2 = \boldsymbol{M} = \boldsymbol{M}_B$. Hence $\boldsymbol{X}$ has an elliptically contoured distribution by Theorem 3.4. See Problem 3.4 for a related result.

Let $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $y \sim \chi^2_d$ be independent. Let $w_i = x_i/(y/d)^{1/2}$ for $i = 1, ..., p$. Then $\boldsymbol{w}$ has a *multivariate t-distribution* with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and degrees of freedom d, an important elliptically contoured distribution. Cornish (1954) showed that the covariance matrix of $\boldsymbol{w}$ is $\text{Cov}(\boldsymbol{w}) = \dfrac{d}{d-2}\boldsymbol{\Sigma}$ for $d > 2$. The case $d = 1$ is known as a multivariate Cauchy distribution. The joint pdf of $\boldsymbol{w}$ is

$$f(\boldsymbol{z}) = \frac{\Gamma((d+p)/2))\ |\boldsymbol{\Sigma}|^{-1/2}}{(\pi d)^{p/2}\Gamma(d/2)}[1 + d^{-1}(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})]^{-(d+p)/2}.$$

See Mardia et al. (1979, pp. 43, 57). See Johnson and Kotz (1972, p. 134) for the special case where the $x_i \sim N(0,1)$.

The following $EC(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution for a $p \times 1$ random vector $\boldsymbol{x}$ is the uniform distribution on a hyperellipsoid where $f(\boldsymbol{z}) = c$ for $\boldsymbol{z}$ in the hyperellipsoid where c is the reciprocal of the volume of the hyperellipsoid. The pdf of the distribution is

$$f(\boldsymbol{z}) = \frac{\Gamma(\frac{p}{2}+1)}{[(p+2)\pi]^{p/2}}|\boldsymbol{\Sigma}|^{-1/2}I[(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}) \le p+2].$$

See Theorem 3.9 where $h^2 = p+2$. Then $E(\boldsymbol{x}) = \boldsymbol{\mu}$ by symmetry and is can be shown that $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}$.

If $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $u_i = \exp(x_i)$ for $i = 1, ..., p$, then $\boldsymbol{u}$ has a multivariate lognormal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. This distribution is not an elliptically contoured distribution. See Problem 3.24.

3.3 The Sample Mean and Sample Covariance Matrix

The population location vector $\boldsymbol{\mu}$ need not be the population mean, but often the population mean is denoted by $\boldsymbol{\mu}$. For elliptically contoured distributions, such as the multivariate normal distribution, $\boldsymbol{\mu}$ is usually the point of symmetry for the population distribution. See Section 3.2. We will now usually use $\boldsymbol{x} = (x_1, ..., x_p)^T$ as a random vector or the observed random vector, depending on the context. Hence $E(\boldsymbol{x}) = (E(x_1), ..., E(x_p))^T$ and $\text{Cov}(\boldsymbol{x}) = (\sigma_{ij}) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T] = E[(\boldsymbol{x} - E(\boldsymbol{x}))\boldsymbol{x}^T] = E(\boldsymbol{x}\boldsymbol{x}^T) - E(\boldsymbol{x})[E(\boldsymbol{x})]^T = \boldsymbol{\Sigma}_{\boldsymbol{x}}$.

Definition 3.7. If the second moments exist, the $p \times p$ population correlation matrix $\text{Cor}(\boldsymbol{x}) = \boldsymbol{\rho}_{\boldsymbol{x}} = (\rho_{ij})$. That is, the ij entry of $\text{Cor}(\boldsymbol{x})$ is $\text{Cor}(X_i, X_j) =$

$$\frac{\sigma_{ij}}{\sigma_i \sigma_j} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}.$$

Let the $p \times p$ population standard deviation matrix

$$\boldsymbol{\Delta} = \text{diag}(\sqrt{\sigma_{11}}, ..., \sqrt{\sigma_{\text{pp}}}).$$

Then

$$\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{\Delta}\boldsymbol{\rho}_{\boldsymbol{x}}\boldsymbol{\Delta}, \tag{3.12}$$

and

$$\boldsymbol{\rho}_{\boldsymbol{x}} = \boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Delta}^{-1}. \tag{3.13}$$

Let the population standardized random variables

$$Z_i = \frac{X_i - E(X_i)}{\sqrt{\sigma_{ii}}}$$

for $i = 1, ..., p$. Then $\text{Cor}(\boldsymbol{x}) = \boldsymbol{\rho}_{\boldsymbol{x}} = \text{Cov}(\boldsymbol{z})$ is the covariance matrix of $\boldsymbol{z} = (Z_1, ..., Z_p)^T$.

Definition 3.8. Let random vectors $\boldsymbol{x}$ be $p \times 1$ and $\boldsymbol{y}$ be $q \times 1$. The *population covariance matrix* of $\boldsymbol{x}$ with $\boldsymbol{y}$ is the $p \times q$ matrix

$$\text{Cov}(\boldsymbol{x}, \boldsymbol{y}) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{y} - E(\boldsymbol{y}))^T] =$$

$$E[(\boldsymbol{x} - E(\boldsymbol{x}))\boldsymbol{y}^T] = E(\boldsymbol{x}\boldsymbol{y}^T) - E(\boldsymbol{x})[E(\boldsymbol{y})]^T = \boldsymbol{\Sigma}_{\boldsymbol{x},\boldsymbol{y}}$$

assuming the expected values exist. Note that the $q \times p$ matrix $\text{Cov}(\boldsymbol{y}, \boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{y},\boldsymbol{x}} = \boldsymbol{\Sigma}_{\boldsymbol{x},\boldsymbol{y}}^T$, and $\text{Cov}(\boldsymbol{x}) = \text{Cov}(\boldsymbol{x}, \boldsymbol{x})$.

Definition 3.9. Let $x_{1j}, ..., x_{nj}$ be measurements on the jth random variable X_j corresponding to the jth column of the data matrix $\boldsymbol{W}$. The

jth *sample mean* is $\overline{x}_j = \frac{1}{n}\sum_{k=1}^{n} x_{kj}$. The *sample covariance* S_{ij} estimates $\mathrm{Cov}(X_i, X_j) = \sigma_{ij}$, and

$$S_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j).$$

$S_{ii} = S_i^2$ is the *sample variance* that estimates the population variance $\sigma_{ii} = \sigma_i^2$. The *sample correlation* r_{ij} estimates the population correlation $\mathrm{Cor}(X_i, X_j) = \rho_{ij}$, and

$$r_{ij} = \frac{S_{ij}}{S_i S_j} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}} = \frac{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)^2}\sqrt{\sum_{k=1}^{n}(x_{kj} - \overline{x}_j)^2}}.$$

Definition 3.10. The **sample mean** or *sample mean vector*

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i = (\overline{x}_1, ..., \overline{x}_p)^T = \frac{1}{n}\boldsymbol{W}^T\mathbf{1}$$

where $\mathbf{1}$ is the $n \times 1$ vector of ones. The **sample covariance matrix**

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T = (S_{ij}).$$

That is, the ij entry of $\boldsymbol{S}$ is the sample covariance S_{ij}. The *classical estimator of multivariate location and dispersion* is $(\overline{\boldsymbol{x}}, \boldsymbol{S})$.

It can be shown that $(n-1)\boldsymbol{S} = \sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T - \overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^T =$

$$\boldsymbol{W}^T\boldsymbol{W} - \frac{1}{n}\boldsymbol{W}^T\mathbf{1}\mathbf{1}^T\boldsymbol{W}.$$

Hence if the *centering matrix* $\boldsymbol{H} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$, then $(n-1)\boldsymbol{S} = \boldsymbol{W}^T\boldsymbol{H}\boldsymbol{W}$.

Definition 3.11. The **sample correlation matrix**

$$\boldsymbol{R} = (r_{ij}).$$

That is, the ij entry of $\boldsymbol{R}$ is the sample correlation r_{ij}.

Let the standardized random variables

$$Z_j = \frac{x_j - \overline{x}_j}{\sqrt{S_{jj}}}$$

for $j = 1, ..., p$. Then the sample correlation matrix $\boldsymbol{R}$ is the sample covariance matrix of the $\boldsymbol{z}_i = (Z_{i1}, ..., Z_{ip})^T$ where $i = 1, ..., n$.

Often it is useful to standardize variables with a robust location estimator and a robust scale estimator. The R function `scale` is useful. The R code below shows how to standardize using

$$Z_j = \frac{x_j - \text{MED}(x_j)}{\text{MAD}(x_j)}$$

for $j = 1, ..., p$. Here $\text{MED}(x_j) = \text{MED}(x_{1j}, ..., x_{nj})$ and $\text{MAD}(x_j) = \text{MAD}(x_{1j}, ..., x_{nj})$ are the sample median and sample median absolute deviation of the data for the jth variable: $x_{1j}, ..., x_{nj}$. See Definitions 2.2 and 2.4. Some of these results are illustrated with the following R code.

```
x <- buxx[,1:3]; cov(x)
                len       nasal     bigonal
len     118299.9257 -191.084603 -104.718925
nasal     -191.0846   18.793905   -1.967121
bigonal   -104.7189   -1.967121   36.796311

cor(x)
                len       nasal     bigonal
len      1.00000000 -0.12815187 -0.05019157
nasal   -0.12815187  1.00000000 -0.07480324
bigonal -0.05019157 -0.07480324  1.00000000
z <- scale(x)
cov(z)
                len       nasal     bigonal
len      1.00000000 -0.12815187 -0.05019157
nasal   -0.12815187  1.00000000 -0.07480324
bigonal -0.05019157 -0.07480324  1.00000000

medd <- apply(x,2,median)
madd <- apply(x,2,mad)/1.4826
z <- scale(x,center=medd,scale=madd)
ddplot4(z)#scaled data still has 5 outliers
cov(z)    #in the length variable
                len      nasal   bigonal
len     4731.997028 -12.738974 -6.981262
nasal    -12.738974   2.088212 -0.218569
bigonal   -6.981262  -0.218569  4.088479

cor(z)
                len       nasal     bigonal
len      1.00000000 -0.12815187 -0.05019157
nasal   -0.12815187  1.00000000 -0.07480324
```

```
bigonal -0.05019157 -0.07480324  1.00000000

apply(z,2,median)
len   nasal bigonal
0       0       0
#scaled data has coord. median = (0,0,0)^T
apply(z,2,mad)/1.4826
len   nasal bigonal
1       1       1 #scaled data has unit MAD
```

Notation. A *rule of thumb* is a rule that often but not always works well in practice.

Rule of thumb 3.1. Multivariate procedures start to give good results for $n \geq 10p$, especially if the distribution is close to multivariate normal. In particular, we want $n \geq 10p$ for the sample covariance and correlation matrices. For procedures with large sample theory on a large class of distributions, for any value of n, there are always distributions where the results will be poor, but will eventually be good for larger sample sizes. Norman and Streiner (1986, pp. 122, 130, 157) gave this rule of thumb and note that some authors recommend $n \geq 30p$. This rule of thumb is much like the rule of thumb that says the central limit theorem normal approximation for $\overline{Y}$ starts to be good for many distributions for $n \geq 30$. See the paragraph below Theorem 11.8.

The population and sample correlation are measures of the strength of a **linear relationship** between two random variables, satisfying $-1 \leq \rho_{ij} \leq 1$ and $-1 \leq r_{ij} \leq 1$. Let the $p \times p$ sample standard deviation matrix

$$\boldsymbol{D} = \text{diag}(\sqrt{\text{S}_{11}}, ..., \sqrt{\text{S}_{\text{pp}}}).$$

Then

$$\boldsymbol{S} = \boldsymbol{DRD}, \tag{3.14}$$

and

$$\boldsymbol{R} = \boldsymbol{D}^{-1}\boldsymbol{S}\boldsymbol{D}^{-1}. \tag{3.15}$$

3.4 Mahalanobis Distances

In the multivariate location and dispersion model, sample Mahalanobis distances play a role similar to that of residuals in multiple linear regression.

Definition 3.12. Let $\boldsymbol{\Sigma}$ be a positive definite symmetric dispersion matrix. Then the *Mahalanobis distance* of $\boldsymbol{x}$ from the vector $\boldsymbol{\mu}$ is

$$D_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sqrt{(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})}.$$

The *population squared Mahalanobis distance*

$$D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}). \tag{3.16}$$

Estimators of multivariate location and dispersion are of interest. Let the observed data $\boldsymbol{x}_i$ for $i = 1, ..., n$ be collected in an $n \times p$ matrix $\boldsymbol{W}$ with n rows $\boldsymbol{x}_1^T, ..., \boldsymbol{x}_n^T$. Let the $p \times 1$ column vector $T(\boldsymbol{W})$ be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C}(\boldsymbol{W})$ be a dispersion estimator. If $(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ then the *sample squared Mahalanobis distance* is

$$D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}).$$

The word "sample" is often suppressed.

Definition 3.13. The ith *squared sample Mahalanobis distance* is

$$D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{x}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T(\boldsymbol{W})) \tag{3.17}$$

for each case $\boldsymbol{x}_i$.

Notice that D_i^2 is a random variable (scalar valued). Notice that the term $\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu})$ is the $p-$dimensional analog to the z-score used to transform a univariate $N(\mu, \sigma^2)$ random variable into a $N(0, 1)$ random variable. Hence the sample Mahalanobis distance $D_i = \sqrt{D_i^2}$ is an analog of the absolute value $|Z_i|$ of the sample Z-score $Z_i = (X_i - \overline{X})/\hat{\sigma}$. Also notice that the Euclidean distance of $\boldsymbol{x}_i$ from the estimate of center $T(\boldsymbol{W})$ is $D_i(T(\boldsymbol{W}), \boldsymbol{I}_p)$ where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix.

Notation: Recall that a square symmetric $p \times p$ matrix $\boldsymbol{A}$ has an *eigenvalue* λ with corresponding *eigenvector* $\boldsymbol{x} \neq \boldsymbol{0}$ if

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}. \tag{3.18}$$

The eigenvalues of $\boldsymbol{A}$ are real since $\boldsymbol{A}$ is symmetric. Note that if constant $c \neq 0$ and $\boldsymbol{x}$ is an eigenvector of $\boldsymbol{A}$, then $c\ \boldsymbol{x}$ is an eigenvector of $\boldsymbol{A}$. Let $\boldsymbol{e}$ be an eigenvector of $\boldsymbol{A}$ with unit length $\|\boldsymbol{e}\| = \sqrt{\boldsymbol{e}^T \boldsymbol{e}} = 1$. Then $\boldsymbol{e}$ and $-\boldsymbol{e}$ are eigenvectors with unit length, and $\boldsymbol{A}$ has p eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), ..., (\lambda_p, \boldsymbol{e}_p)$. Since $\boldsymbol{A}$ is symmetric, the eigenvectors are chosen such that the $\boldsymbol{e}_i$ are orthogonal: $\boldsymbol{e}_i^T \boldsymbol{e}_j = 0$ for $i \neq j$. The symmetric matrix $\boldsymbol{A}$ is positive definite iff all of its eigenvalues are positive, and positive semidefinite iff all of its eigenvalues are nonnegative. If $\boldsymbol{A}$ is positive semidefinite, let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. If $\boldsymbol{A}$ is positive definite, then $\lambda_p > 0$.

Theorem 3.7. Let $\boldsymbol{A}$ be a $p \times p$ symmetric matrix with eigenvector eigenvalue pairs $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), ..., (\lambda_p, \boldsymbol{e}_p)$ where $\boldsymbol{e}_i^T \boldsymbol{e}_i = 1$ and $\boldsymbol{e}_i^T \boldsymbol{e}_j = 0$ if $i \neq j$ for $i = 1, ..., p$. Then the *spectral decomposition* of $\boldsymbol{A}$ is

$$\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T = \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1^T + \cdots + \lambda_p \boldsymbol{e}_p \boldsymbol{e}_p^T.$$

Using the same notation as Johnson and Wichern (1988, pp. 50-51), let $\boldsymbol{P} = [\boldsymbol{e}_1 \ \boldsymbol{e}_2 \ \cdots \ \boldsymbol{e}_p]$ be the $p \times p$ orthogonal matrix with ith column $\boldsymbol{e}_i$. Then $\boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}$. Let $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_p)$ and let $\boldsymbol{\Lambda}^{1/2} = \text{diag}(\sqrt{\lambda_1}, ..., \sqrt{\lambda_p})$. If $\boldsymbol{A}$ is a positive definite $p \times p$ symmetric matrix with spectral decomposition $\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T$, then $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^T$ and

$$\boldsymbol{A}^{-1} = \boldsymbol{P}\boldsymbol{\Lambda}^{-1}\boldsymbol{P}^T = \sum_{i=1}^{p} \frac{1}{\lambda_i} \boldsymbol{e}_i \boldsymbol{e}_i^T.$$

Theorem 3.8. Let $\boldsymbol{A}$ be a positive definite $p \times p$ symmetric matrix with spectral decomposition $\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T$. The *square root matrix* $\boldsymbol{A}^{1/2} = \boldsymbol{P}\boldsymbol{\Lambda}^{1/2}\boldsymbol{P}^T$ is a positive definite symmetric matrix such that $\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2} = \boldsymbol{A}$.

Points $\boldsymbol{x}$ with the same distance $D_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ lie on a hyperellipsoid where the shape of the hyperellipsoid is determined by the eigenvectors and eigenvalues of $\boldsymbol{\Sigma}$: $(\lambda_1, \boldsymbol{e}_1), ..., (\lambda_p, \boldsymbol{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. Note $\boldsymbol{\Sigma}^{-1}$ has the same eigenvectors as $\boldsymbol{\Sigma}$ but eigenvalues equal to $1/\lambda_i$ since $\boldsymbol{\Sigma}\boldsymbol{e} = \lambda\boldsymbol{e}$ iff $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{e} = \boldsymbol{e} = \boldsymbol{\Sigma}^{-1}\lambda\boldsymbol{e}$. Then divide both sides by $\lambda > 0$ since $\boldsymbol{\Sigma} > 0$ and is symmetric. Let $\boldsymbol{w} = \boldsymbol{x} - \boldsymbol{\mu}$. Then points at squared distance $\boldsymbol{w}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{w} = h^2$ from the origin lie on the hyperellipsoid centered at the origin whose axes are given by the eigenvectors of $\boldsymbol{\Sigma}$ where the half length in the direction of $\boldsymbol{e}_i$ is $h\sqrt{\lambda_i}$.

Theorem 3.9. Let $\boldsymbol{\Sigma}$ be a positive definite symmetric matrix, e.g. a dispersion matrix. Let $U = D_{\boldsymbol{x}}^2 = D_{\boldsymbol{x}}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The hyperellipsoid

$$\{\boldsymbol{x} | D_{\boldsymbol{x}}^2 \leq h^2\} = \{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \leq h^2\},$$

where $h^2 = u_{1-\alpha}$ and $P(U \leq u_{1-\alpha}) = 1 - \alpha$, is the highest density region covering $1 - \alpha$ of the mass for an elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution (see Definitions 3.5 and 3.6) if g is continuous and decreasing. Let $\boldsymbol{w} = \boldsymbol{x} - \boldsymbol{\mu}$. Then points at a squared distance $\boldsymbol{w}^T\boldsymbol{S}^{-1}\boldsymbol{w} = h^2$ from the origin lie on the hyperellipsoid centered at the origin whose axes are given by the eigenvectors $\boldsymbol{e}_i$ where the half length in the direction of $\boldsymbol{e}_i$ is $h\sqrt{\lambda_i}$. The volume of the hyperellipsoid is

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} |\boldsymbol{\Sigma}|^{1/2} h^p.$$

Theorem 3.10. Let the symmetric sample covariance matrix $\boldsymbol{S}$ be positive definite with eigenvalue eigenvector pairs $(\hat{\lambda}_i, \hat{\boldsymbol{e}}_i)$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p > 0$. The hyperellipsoid

$$\{\boldsymbol{x} | D^2_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq h^2\} = \{\boldsymbol{x} : (\boldsymbol{x} - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x} - \overline{\boldsymbol{x}}) \leq h^2\}$$

is centered at $\overline{\boldsymbol{x}}$. The volume of the hyperellipsoid is

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} |\boldsymbol{S}|^{1/2} h^p.$$

Let $\boldsymbol{w} = \boldsymbol{x} - \overline{\boldsymbol{x}}$. Then points at a squared distance $\boldsymbol{w}^T \boldsymbol{S}^{-1} \boldsymbol{w} = h^2$ from the origin lie on the hyperellipsoid centered at the origin whose axes are given by the eigenvectors $\hat{\boldsymbol{e}}_i$ where the half length in the direction of $\hat{\boldsymbol{e}}_i$ is $h\sqrt{\hat{\lambda}_i}$.

From Theorem 3.9, the volume of the hyperellipsoid $\{\boldsymbol{x} | D^2_{\boldsymbol{x}} \leq h^2\}$ is proportional to $|\boldsymbol{S}|^{1/2}$ so the squared volume is proportional to $|\boldsymbol{S}|$. Large $|\boldsymbol{S}|$ corresponds to large volume while small $|\boldsymbol{S}|$ corresponds to small volume.

Definition 3.14. The *generalized sample variance* $= |\boldsymbol{S}| = \det(\boldsymbol{S})$.

Following Johnson and Wichern (1988, pp. 103-106), a generalized variance of zero is indicative of extreme degeneracy, and $|\boldsymbol{S}| = 0$ implies that at least one variable X_i is not needed given the other $p - 1$ variables are in the multivariate model. Two necessary conditions for $|\boldsymbol{S}| \neq 0$ are $n > p$ and that $\boldsymbol{S}$ has full rank p. If $\boldsymbol{1}$ is an $n \times 1$ vector of ones, then

$$(n-1)\boldsymbol{S} = (\boldsymbol{W} - \boldsymbol{1}\overline{\boldsymbol{x}}^T)^T (\boldsymbol{W} - \boldsymbol{1}\overline{\boldsymbol{x}}^T),$$

and $\boldsymbol{S}$ is of full rank p iff $\boldsymbol{W} - \boldsymbol{1}\overline{\boldsymbol{x}}^T$ is of full rank p.

If $\boldsymbol{X}$ and $\boldsymbol{Z}$ have dispersion matrices $\boldsymbol{\Sigma}$ and $c\boldsymbol{\Sigma}$ where $c > 0$, then the dispersion matrices have the same shape. The dispersion matrices determine the shape of the hyperellipsoid $\{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \leq h^2\}$. Figure 3.1 was made with the *Arc* software of Cook and Weisberg (1999a). The 10%, 30%, 50%, 70%, 90%, and 98% highest density regions are shown for two multivariate normal (MVN) distributions. Both distributions have $\boldsymbol{\mu} = \boldsymbol{0}$. In Figure 3.1a),

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 4 \end{pmatrix}.$$

Note that the ellipsoids are narrow with high positive correlation. In Figure 3.1b),

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & -0.4 \\ -0.4 & 1 \end{pmatrix}.$$

Note that the ellipsoids are wide with negative correlation. The highest density ellipsoids are superimposed on a scatterplot of a sample of size 100 from each distribution.

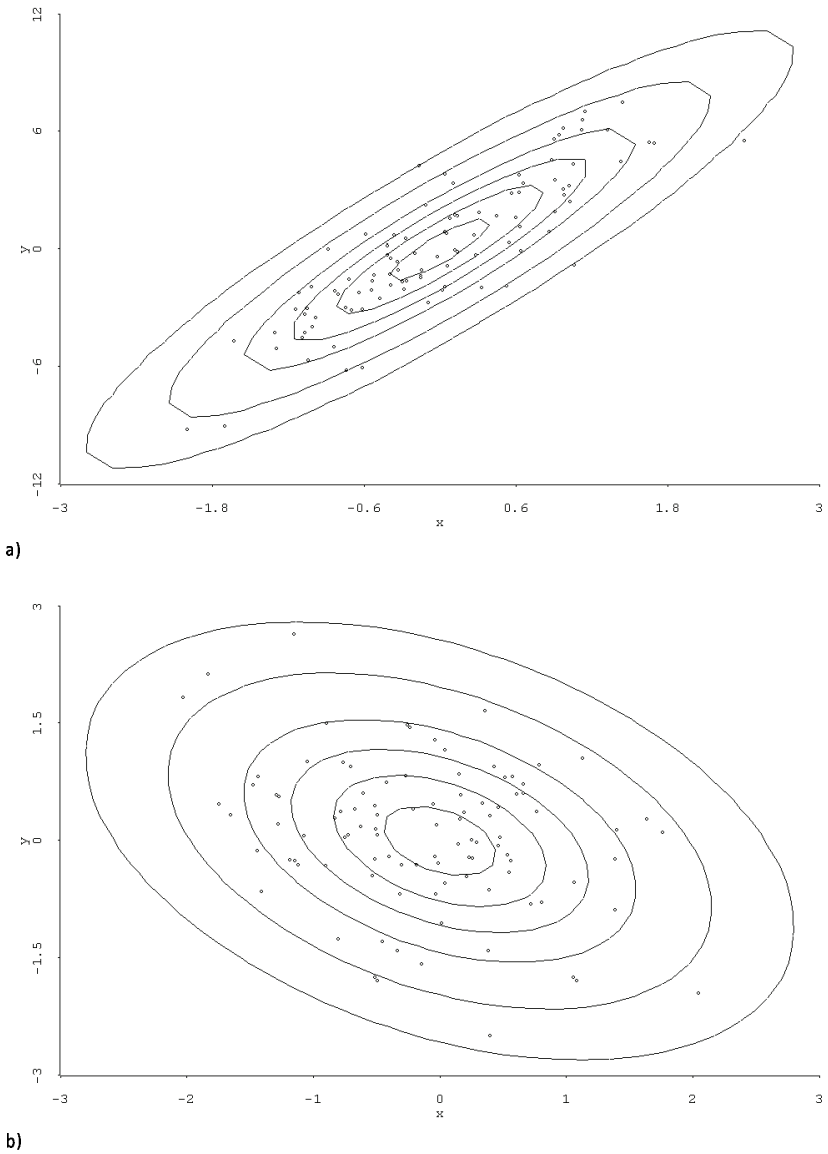

Fig. 3.1 Highest Density Regions for 2 MVN Distributions

Example 3.3. The contours of constant density for the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution are ellipsoids defined by $\boldsymbol{x}$ such that $(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = a^2$. An $\alpha-$density region R_α is a set such that $P(\boldsymbol{X} \in R_\alpha) = \alpha$, and for the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution, the regions of highest density are sets of the form

$$\{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \leq \chi^2_p(\alpha)\} = \{\boldsymbol{x} : D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leq \chi^2_p(\alpha)\}$$

where $P(W \leq \chi^2_p(\alpha)) = \alpha$ if $W \sim \chi^2_p$. If the $\boldsymbol{X}_i$ are n iid random vectors each with a $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ pdf, then a scatterplot of $X_{i,k}$ versus $X_{i,j}$ should be ellipsoidal for $k \neq j$. Similar statements hold if $\boldsymbol{X}$ is $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$, but the α-density region will use a constant U_α obtained from Equation (3.10).

3.5 Equivariance and Breakdown

Equivariance and breakdown properties are very weak compared to properties like consistency, but will be useful for the theory of practical robust MLD estimators. Before defining an important equivariance property, some notation is needed. Again assume that the data is collected in an $n \times p$ data matrix $\boldsymbol{W}$. Let $\boldsymbol{B} = \boldsymbol{1}\boldsymbol{b}^T$ where $\boldsymbol{1}$ is an $n \times 1$ vector of ones and $\boldsymbol{b}$ is a $p \times 1$ constant vector. Hence the ith row of $\boldsymbol{B}$ is $\boldsymbol{b}_i^T \equiv \boldsymbol{b}^T$ for $i = 1, ..., n$. For such a matrix $\boldsymbol{B}$, consider the affine transformation $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}$ where $\boldsymbol{A}$ is any nonsingular $p \times p$ matrix. An affine transformation changes $\boldsymbol{x}_i$ to $\boldsymbol{z}_i = \boldsymbol{A}\boldsymbol{x}_i + \boldsymbol{b}$ for $i = 1, ..., n$, and affine equivariant multivariate location and dispersion estimators change in natural ways.

Definition 3.15. The multivariate location and dispersion estimator $(T, \boldsymbol{C})$ is *affine equivariant* if

$$T(\boldsymbol{Z}) = T(\boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}) = \boldsymbol{A}T(\boldsymbol{W}) + \boldsymbol{b}, \tag{3.19}$$

$$\text{and } \boldsymbol{C}(\boldsymbol{Z}) = \boldsymbol{C}(\boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}) = \boldsymbol{A}\boldsymbol{C}(\boldsymbol{W})\boldsymbol{A}^T. \tag{3.20}$$

The following theorem shows that the Mahalanobis distances are invariant under affine transformations. See Rousseeuw and Leroy (1987, pp. 252-262) for similar results. Thus if $(T, \boldsymbol{C})$ is affine equivariant, so is $(T, D^2_{(c_n)}(T, \boldsymbol{C})\ \boldsymbol{C})$ where $D^2_{(j)}(T, \boldsymbol{C})$ is the jth order statistic of the D^2_i.

Theorem 3.11. If $(T, \boldsymbol{C})$ is affine equivariant, then

$$D^2_i(\boldsymbol{W}) \equiv D^2_i(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = D^2_i(T(\boldsymbol{Z}), \boldsymbol{C}(\boldsymbol{Z})) \equiv D^2_i(\boldsymbol{Z}). \tag{3.21}$$

Proof. Since $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}$ has ith row $\boldsymbol{z}_i^T = \boldsymbol{x}_i^T\boldsymbol{A}^T + \boldsymbol{b}^T$,

$$D^2_i(\boldsymbol{Z}) = [\boldsymbol{z}_i - T(\boldsymbol{Z})]^T \boldsymbol{C}^{-1}(\boldsymbol{Z})[\boldsymbol{z}_i - T(\boldsymbol{Z})]$$

$$= [\boldsymbol{A}(\boldsymbol{x}_i - T(\boldsymbol{W}))]^T [\boldsymbol{A}\boldsymbol{C}(\boldsymbol{W})\boldsymbol{A}^T]^{-1} [\boldsymbol{A}(\boldsymbol{x}_i - T(\boldsymbol{W}))]$$
$$= [\boldsymbol{x}_i - T(\boldsymbol{W})]^T \boldsymbol{C}^{-1}(\boldsymbol{W})[\boldsymbol{x}_i - T(\boldsymbol{W})] = D_i^2(\boldsymbol{W}). \ \square$$

Warning: Estimators that use randomly chosen elemental sets or projections are not affine equivariant since these estimators often change when they are computed several times (corresponding to the identity transformation $\boldsymbol{A} = \boldsymbol{I}_p$). Such estimators can sometimes be made pseudo-affine equivariant by using the same fixed random number seed and random number generator each time the estimator is used. Then the pseudo-affine equivariance of the estimator depends on the random number seed and the random number generator, and such estimators are not as attractive as affine equivariant estimators that do not depend on a fixed random number seed and random number generator.

Next, a standard definition of breakdown is given for estimators of multivariate location and dispersion. The following notation will be useful. Let $\boldsymbol{W}$ denote the $n \times p$ data matrix with ith row $\boldsymbol{x}_i^T$ corresponding to the ith case. Let $\boldsymbol{w}_1, ...\boldsymbol{w}_n$ be the contaminated data after d_n of the $\boldsymbol{x}_i$ have been replaced by arbitrarily bad contaminated cases. Let $\boldsymbol{W}_d^n$ denote the $n \times p$ data matrix with ith row $\boldsymbol{w}_i^T$. Then the contamination fraction is $\gamma_n = d_n/n$. Let $(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W}))$ denote an estimator of multivariate location and dispersion where the $p \times 1$ vector $T(\boldsymbol{W})$ is an estimator of location and the $p \times p$ symmetric positive semidefinite matrix $\boldsymbol{C}(\boldsymbol{W})$ is an estimator of dispersion. A theorem from multivariate analysis shows that if $\boldsymbol{C}(\boldsymbol{W}_d^n) > 0$, then $\max_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^T \boldsymbol{C}(\boldsymbol{W}_d^n)\boldsymbol{a} = \lambda_1$ and $\min_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^T \boldsymbol{C}(\boldsymbol{W}_d^n)\boldsymbol{a} = \lambda_p$. See Olive (2017b, p. 7) and Johnson and Wichern (1988, pp, 64-65, 184). A high breakdown dispersion estimator $\boldsymbol{C}$ is positive definite if the amount of contamination is less than the breakdown value. Since $\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a} = \sum_{i=1}^p \sum_{j=1}^p c_{ij} a_i a_j$, the largest eigenvalue λ_1 is bounded as $\boldsymbol{W}_d^n$ varies iff $\boldsymbol{C}(\boldsymbol{W}_d^n)$ is bounded as $\boldsymbol{W}_d^n$ varies.

Definition 3.16. The *breakdown value* of the multivariate location estimator T at $\boldsymbol{W}$ is

$$B(T, \boldsymbol{W}) = \min \left\{ \frac{d_n}{n} : \sup_{\boldsymbol{W}_d^n} \|T(\boldsymbol{W}_d^n)\| = \infty \right\}$$

where the supremum is over all possible corrupted samples $\boldsymbol{W}_d^n$ and $1 \leq d_n \leq n$. Let $\lambda_1(\boldsymbol{C}(\boldsymbol{W})) \geq \cdots \geq \lambda_p(\boldsymbol{C}(\boldsymbol{W})) \geq 0$ denote the eigenvalues of the dispersion estimator applied to data $\boldsymbol{W}$. The estimator $\boldsymbol{C}$ breaks down if the smallest eigenvalue can be driven to zero or if the largest eigenvalue can be driven to ∞. Hence the *breakdown value* of the dispersion estimator is

$$B(\boldsymbol{C}, \boldsymbol{W}) = \min \left\{ \frac{d_n}{n} : \sup_{\boldsymbol{W}_d^n} \max \left[\frac{1}{\lambda_p(\boldsymbol{C}(\boldsymbol{W}_d^n))}, \lambda_1(\boldsymbol{C}(\boldsymbol{W}_d^n)) \right] = \infty \right\}.$$

Definition 3.17. Let γ_n be the breakdown value of $(T, \boldsymbol{C})$. *High breakdown (HB) statistics* have $\gamma_n \to 0.5$ as $n \to \infty$ if the (uncontaminated) clean data are in *general position*: no more than p points of the clean data lie on any $(p-1)$-dimensional hyperplane. Estimators are *zero breakdown* if $\gamma_n \to 0$ and *positive breakdown* if $\gamma_n \to \gamma > 0$ as $n \to \infty$.

Note that if the number of outliers is less than the number needed to cause breakdown, then $\|T\|$ is bounded and the eigenvalues are bounded away from 0 and ∞. Also, the bounds do not depend on the outliers but do depend on the estimator $(T, \boldsymbol{C})$ and on the clean data $\boldsymbol{W}$.

The following result shows that a multivariate location estimator T basically "breaks down" if the d outliers can make the median Euclidean distance $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|)$ arbitrarily large where $\boldsymbol{w}_i^T$ is the ith row of $\boldsymbol{W}_d^n$. Thus a multivariate location estimator T will not break down if T can not be driven out of some ball of (possibly huge) radius r about the origin. For an affine equivariant estimator, the largest possible breakdown value is $n/2$ or $(n+1)/2$ for n even or odd, respectively. Hence in the proof of the following result, we could replace $d_n < d_T$ by $d_n < \min(n/2, d_T)$.

Theorem 3.12. Fix n. If nonequivariant estimators (that may have a breakdown value of greater than 1/2) are excluded, then a multivariate location estimator has a breakdown value of d_T/n iff $d_T = d_{T,n}$ is the smallest number of arbitrarily bad cases that can make the median Euclidean distance $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|)$ arbitrarily large.

Proof. Suppose the multivariate location estimator T satisfies $\|T(\boldsymbol{W}_d^n)\| \leq M$ for some constant M if $d_n < d_T$. Note that for a fixed data set $\boldsymbol{W}_d^n$ with ith row $\boldsymbol{w}_i$, the median Euclidean distance $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|) \leq \max_{i=1,\ldots,n} \|\boldsymbol{x}_i - T(\boldsymbol{W}_d^n)\| \leq \max_{i=1,\ldots,n} \|\boldsymbol{x}_i\| + M$ if $d_n < d_T$. Similarly, suppose $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|) \leq M$ for some constant M if $d_n < d_T$, then $\|T(\boldsymbol{W}_d^n)\|$ is bounded if $d_n < d_T$. □

Since the coordinatewise median $\text{MED}(\boldsymbol{W})$ is a HB estimator of multivariate location, it is also true that a multivariate location estimator T will not break down if T can not be driven out of some ball of radius r about $\text{MED}(\boldsymbol{W})$. Hence $(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ is a HB estimator of MLD.

If a high breakdown estimator $(T, \boldsymbol{C}) \equiv (T(\boldsymbol{W}_d^n), \boldsymbol{C}(\boldsymbol{W}_d^n))$ is evaluated on the contaminated data $\boldsymbol{W}_d^n$, then the location estimator T is contained in some ball about the origin of radius r, and $0 < a < \lambda_p \leq \lambda_1 < b$ where the constants a, r, and b depend on the clean data and $(T, \boldsymbol{C})$, but not on $\boldsymbol{W}_d^n$ if the number of outliers d_n satisfies $0 \leq d_n < n\gamma_n < n/2$ where the breakdown value $\gamma_n \to 0.5$ as $n \to \infty$.

The following theorem will be used to show that if the classical estimator $(\overline{\boldsymbol{X}}_B, \boldsymbol{S}_B)$ is applied to $c_n \approx n/2$ cases contained in a ball about the origin of radius r where r depends on the clean data but not on $\boldsymbol{W}_d^n$, then $(\overline{\boldsymbol{X}}_B, \boldsymbol{S}_B)$ is a high breakdown estimator.

Theorem 3.13. If the classical estimator $(\overline{\boldsymbol{X}}_B, \boldsymbol{S}_B)$ is applied to c_n cases that are contained in some bounded region where $p+1 \leq c_n \leq n$, then the maximum eigenvalue λ_1 of $\boldsymbol{S}_B$ is bounded.

Proof. The largest eigenvalue of a $p \times p$ matrix $\boldsymbol{A}$ is bounded above by $p \max |a_{i,j}|$ where $a_{i,j}$ is the (i,j) entry of $\boldsymbol{A}$. See Datta (1995, p. 403). Denote the c_n cases by $\boldsymbol{z}_1, ..., \boldsymbol{z}_{c_n}$. Then the (i,j)th element $a_{i,j}$ of $\boldsymbol{A} = \boldsymbol{S}_B$ is

$$a_{i,j} = \frac{1}{c_n - 1} \sum_{m=1}^{c_n} (z_{i,m} - \overline{z}_i)(z_{j,m} - \overline{z}_j).$$

Hence the maximum eigenvalue λ_1 is bounded. □

The determinant $det(\boldsymbol{S}) = |\boldsymbol{S}|$ of $\boldsymbol{S}$ is known as the *generalized sample variance*. See Definition 3.14. Consider the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq D^2_{(c_n)}\} \tag{3.22}$$

where $D^2_{(c_n)}$ is the c_nth smallest squared Mahalanobis distance based on $(T, \boldsymbol{C})$. This hyperellipsoid contains the c_n cases with the smallest D^2_i. 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 where $b > 0$. The classical, RFCH, and RMVN estimators satisfy this assumption. For $h > 0$, the hyperellipsoid

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

has volume equal to

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

If $h^2 = D^2_{(c_n)}$, then the volume is proportional to the square root of the determinant $|\boldsymbol{S}_M|^{1/2}$, and this volume will be positive unless extreme degeneracy is present among the c_n cases. See Johnson and Wichern (1988, pp. 103-104).

3.6 The Concentration Algorithm

Concentration algorithms are widely used since impractical brand name estimators, such as the MCD estimator given in Definition 3.18, take too long to compute. The concentration algorithm, defined in Definition 3.19, uses K starts and attractors. A *start* is an initial estimator, and an *attractor* is an estimator obtained by refining the start. For example, let the start be the classical estimator $(\overline{\boldsymbol{x}}, \boldsymbol{S})$. Then the attractor could be the classical estimator $(T_1, \boldsymbol{C}_1)$ applied to the half set of cases with the smallest Mahalanobis

distances. This concentration algorithm uses one concentration step, but the process could be iterated for k concentration steps, producing an estimator $(T_k, \boldsymbol{C}_k)$

If more than one attractor is used, then some criterion is needed to select which of the K attractors is to be used in the final estimator. If each attractor $(T_{k,j}, \boldsymbol{C}_{k,j})$ is the classical estimator applied to $c_n \approx n/2$ cases, then the minimum covariance determinant (MCD) criterion is often used: choose the attractor that has the minimum value of $det(\boldsymbol{C}_{k,j})$ where $j = 1, ..., K$.

This chapter will explain the concentration algorithm, explain why the MCD criterion is useful but can be improved, provide some theory for practical robust multivariate location and dispersion estimators, and show how the set of cases used to compute the recommended RMVN or RFCH estimator can be used to create robust multivariate analogs of methods such as principal component analysis and canonical correlation analysis. The RMVN and RFCH estimators are reweighted versions of the practical FCH estimator, given in Definition 3.22.

Definition 3.18. Consider the subset J_o of $c_n \approx n/2$ observations whose sample covariance matrix has the lowest determinant among all $C(n, c_n)$ subsets of size c_n. Let T_{MCD} and $\boldsymbol{C}_{MCD}$ denote the sample mean and sample covariance matrix of the c_n cases in J_o. Then the *minimum covariance determinant* MCD(c_n) estimator is $(T_{MCD}(\boldsymbol{W}), \boldsymbol{C}_{MCD}(\boldsymbol{W}))$.

Here

$$C(n, i) = \binom{n}{i} = \frac{n!}{i! \;\; (n-i)!}$$

is the binomial coefficient.

Remark 3.3. Note that for fixed h, the MCD estimator corresponds to the sample mean and covariance estimator of c_n cases such that the hyperellipsoid of Theorem 3.10 has the smallest volume.

The MCD estimator is a high breakdown (HB) estimator, and the value $c_n = \lfloor (n + p + 1)/2 \rfloor$ is often used as the default. The MCD estimator is the pair

$$(\hat{\beta}_{LTS}, Q_{LTS}(\hat{\beta}_{LTS})/(c_n - 1))$$

in the location model where LTS stands for the least trimmed sum of squares estimator. See Section 2.12 and Chapter 5. The population analog of the MCD estimator is closely related to the hyperellipsoid of highest concentration that contains $c_n/n \approx$ half of the mass. The MCD estimator is a $\sqrt{n}$ consistent HB asymptotically normal estimator for $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ where a_{MCD} is some positive constant when the data $\boldsymbol{x}_i$ are iid from a large class of distributions. See Cator and Lopuhaä (2010, 2012) who extended some results of Butler et al. (1993).

Computing robust covariance estimators can be very expensive. For example, to compute the exact MCD(c_n) estimator (T_{MCD}, C_{MCD}), we need to

consider the $C(n, c_n)$ subsets of size c_n. Woodruff and Rocke (1994, p. 893) noted that if 1 billion subsets of size 101 could be evaluated per second, it would require 10^{33} millenia to search through all $C(200, 101)$ subsets if the sample size $n = 200$.

Hence algorithm estimators will be used to approximate the robust estimators. Elemental sets are the key ingredient for both *basic resampling* and *concentration* algorithms.

Definition 3.19. Suppose that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are $p \times 1$ vectors of observed data. For the multivariate location and dispersion model, an *elemental set* J is a set of $p + 1$ cases. An elemental start is the sample mean and sample covariance matrix of the data corresponding to J. In a *concentration algorithm,* let $(T_{-1,j}, \boldsymbol{C}_{-1,j})$ be the jth start (not necessarily elemental) and compute all n Mahalanobis distances $D_i(T_{-1,j}, \boldsymbol{C}_{-1,j})$. At the next iteration, the classical estimator $(T_{0,j}, \boldsymbol{C}_{0,j}) = (\overline{\boldsymbol{x}}_{0,j}, \boldsymbol{S}_{0,j})$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest distances. This iteration can be continued for k *concentration steps* resulting in the sequence of estimators $(T_{-1,j}, \boldsymbol{C}_{-1,j}), (T_{0,j}, \boldsymbol{C}_{0,j}), ..., (T_{k,j}, \boldsymbol{C}_{k,j})$. The result of the iteration $(T_{k,j}, \boldsymbol{C}_{k,j})$ is called the jth *attractor*. If K_n starts are used, then $j = 1, ..., K_n$. The *concentration attractor*, $(T_A, \boldsymbol{C}_A)$, is the attractor chosen by the algorithm. The attractor is used to obtain the final estimator. A common choice is the attractor that has the smallest determinant $det(\boldsymbol{C}_{k,j})$. The *basic resampling algorithm* estimator is a special case where $k = -1$ or $k = 0$ so that the attractor is the start: $(\overline{\boldsymbol{x}}_{k,j}, \boldsymbol{S}_{k,j}) = (\overline{\boldsymbol{x}}_{-1,j}, \boldsymbol{S}_{-1,j})$, or $(\overline{\boldsymbol{x}}_{k,j}, \boldsymbol{S}_{k,j}) = (\overline{\boldsymbol{x}}_{0,j}, \boldsymbol{S}_{0,j})$. The *elemental basic resampling* estimator uses K_n elemental starts and $k = 0$.

This concentration algorithm is a simplified version of the algorithms given by Rousseeuw and Van Driessen (1999) and Hawkins and Olive (1999a). Using $k = 10$ concentration steps often works well. The following theorem is useful and shows that $det(\boldsymbol{S}_{0,j})$ tends to be greater than the determinant of the attractor $det(\boldsymbol{S}_{k,j})$.

Theorem 3.14: Rousseeuw and Van Driessen (1999, p. 214). Suppose that the classical estimator $(\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$ is computed from c_n cases and that the n Mahalanobis distances $D_i \equiv D_i(\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$ are computed. If $(\overline{\boldsymbol{x}}_{t+1,j}, \boldsymbol{S}_{t+1,j})$ is the classical estimator computed from the c_n cases with the smallest Mahalanobis distances D_i, then $det(\boldsymbol{S}_{t+1,j}) \leq det(\boldsymbol{S}_{t,j})$ with equality iff $(\overline{\boldsymbol{x}}_{t+1,j}, \boldsymbol{S}_{t+1,j}) = (\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$.

Starts that use a consistent initial estimator could be used. K_n is the number of starts and k is the number of concentration steps used in the algorithm. Suppose the algorithm estimator uses some criterion to choose an attractor as the final estimator where there are K attractors and K is fixed, e.g. $K = 500$, so K does not depend on n. A crucial observation is that the

theory of the algorithm estimator depends on the theory of the attractors, not on the estimator corresponding to the criterion.

For example, let $(\mathbf{0}, \boldsymbol{I}_p)$ and $(\mathbf{1}, diag(1, 3, ..., p))$ be the high breakdown attractors where $\mathbf{0}$ and $\mathbf{1}$ are the $p \times 1$ vectors of zeroes and ones. If the minimum determinant criterion is used, then the final estimator is $(\mathbf{0}, \boldsymbol{I}_p)$. Although the MCD criterion is used, the algorithm estimator does not have the same properties as the MCD estimator.

Hawkins and Olive (2002) showed that if K randomly selected elemental starts are used with concentration to produce the attractors, then the resulting estimator is inconsistent and zero breakdown if K and k are fixed and free of n. Note that each elemental start can be made to breakdown by changing one case. Hence the breakdown value of the final estimator is bounded by $K/n \to 0$ as $n \to \infty$. Note that the classical estimator computed from h_n randomly drawn cases is an inconsistent estimator unless $h_n \to \infty$ as $n \to \infty$. Thus the classical estimator applied to a randomly drawn elemental set of $h_n = h \equiv p + 1$ cases is an inconsistent estimator, so the K starts and the K attractors are inconsistent.

Theorem 3.15: a) The elemental basic resampling algorithm estimators are inconsistent. b) The elemental concentration and elemental basic resampling algorithm estimators are zero breakdown.

Proof: a) Note that you can not get a consistent estimator by using Kh randomly selected cases since the number of cases Kh needs to go to ∞ for consistency except in degenerate situations.

b) Contaminating all Kh cases in the K elemental sets shows that the breakdown value is bounded by $Kh/n \to 0$, so the estimator is zero breakdown. □

Theorem 3.15 shows that the elemental basic resampling PROGRESS estimators of Rousseeuw (1984), Rousseeuw and Leroy (1987), and Rousseeuw and van Zomeren (1990) with $K = 3000$ are zero breakdown and inconsistent. The Maronna et al. (2006, pp. 198-199) estimators that use $K = 500$ elemental starts and one concentration step ($k = 0$) are inconsistent and zero breakdown. Yohai's two stage estimators need initial consistent high breakdown estimators, such as MCD, but were implemented with the inconsistent zero breakdown elemental basic resampling estimators such as FMCD. See Hawkins and Olive (2002, p. 157). Theorem 5.13 and Remark 5.5 give similar results for multiple linear regression.

The following theorem is useful because it does not depend on the criterion used to choose the attractor. If the algorithm needs to use many attractors to achieve outlier resistance, then the individual attractors have little outlier resistance. Such estimators include elemental concentration algorithms, heuristic and genetic algorithms, and projection algorithms that use randomly chosen projections. Algorithms where all K of the attractors are inconsistent, such as elemental concentration algorithms that use k concentration steps, are especially untrustworthy. You can get consistent estimators if

$K = K_n \to \infty$ or $h = h_n \to \infty$ as $n \to \infty$. You can get high breakdown estimators and avoid singular starts if all $K = K_n = C(n,h)$ elemental sets are used, but such an estimator is impractical.

Remark 3.4. It is unknown whether iterating to convergence, so k is not fixed, results in a consistent or inconsistent estimator. Iteration to convergence does seem to be fairly fast.

Suppose there are K consistent estimators $(T_j, \boldsymbol{C}_j)$ of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ for some constant $a > 0$, each with the same rate n^δ. If $(T_A, \boldsymbol{C}_A)$ is an estimator obtained by choosing one of the K estimators, then $(T_A, \boldsymbol{C}_A)$ is a consistent estimator of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ with rate n^δ by Pratt (1959). See Theorem 11.16.

Theorem 3.16. Suppose the algorithm estimator chooses an attractor as the final estimator where there are K attractors and K is fixed.

i) If all of the attractors are consistent estimators of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$, then the algorithm estimator is a consistent estimator of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$.

ii) If all of the attractors are consistent estimators of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ with the same rate, e.g. n^δ where $0 < \delta \leq 0.5$, then the algorithm estimator is a consistent estimator of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ with the same rate as the attractors.

iii) If all of the attractors are high breakdown, then the algorithm estimator is high breakdown.

iv) Suppose the data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid and $P(\boldsymbol{x}_i = \boldsymbol{\mu}) < 1$. The elemental basic resampling algorithm estimator is inconsistent.

v) The elemental concentration algorithm is zero breakdown.

Proof. i) Choosing from K consistent estimators for $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ results in a consistent estimator for of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$, and ii) follows from Pratt (1959). iii) Let $\gamma_{n,i}$ be the breakdown value of the ith attractor if the clean data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are in general position. The breakdown value γ_n of the algorithm estimator can be no lower than that of the worst attractor: $\gamma_n \geq \min(\gamma_{n,1}, ..., \gamma_{n,K}) \to 0.5$ as $n \to \infty$.

iv) Let $(\overline{\boldsymbol{x}}_{-1,j}, \boldsymbol{S}_{-1,j})$ be the classical estimator applied to a randomly drawn elemental set. Then $\overline{\boldsymbol{x}}_{-1,j}$ is the sample mean applied to $p+1$ iid cases. Hence $E(\boldsymbol{S}_j) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$, $E[\overline{\boldsymbol{x}}_{-1,j}] = E(\boldsymbol{x}) = \boldsymbol{\mu}$, and $\mathrm{Cov}(\overline{\boldsymbol{x}}_{-1,j}) = \mathrm{Cov}(\boldsymbol{x})/(p+1) = \boldsymbol{\Sigma}_{\boldsymbol{x}}/(p+1)$ assuming second moments. So the $(\overline{\boldsymbol{x}}_{-1,j}, \boldsymbol{S}_{-1,j})$ are identically distributed and inconsistent estimators of $(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$. Even without second moments, there exists $\epsilon > 0$ such that $P(\|\overline{\boldsymbol{x}}_{-1,j} - \boldsymbol{\mu}\| > \epsilon) = \delta_\epsilon > 0$ where the probability, ϵ, and δ_ϵ do not depend on n since the distribution of $\overline{\boldsymbol{x}}_{-1,j}$ only depends on the distribution of the iid $\boldsymbol{x}_i$, not on n. Then $P(\min_j \|\overline{\boldsymbol{x}}_{-1,j} - \boldsymbol{\mu}\| > \epsilon) = P(\text{all } \|\overline{\boldsymbol{x}}_{-1,\mathrm{j}} - \boldsymbol{\mu}\| > \epsilon) \to \delta_\epsilon^{\mathrm{K}} > 0$ as $n \to \infty$ where equality would hold if the $\overline{\boldsymbol{x}}_{-1,j}$ were iid. Hence the "best start" that minimizes $\|\overline{\boldsymbol{x}}_{-1,j} - \boldsymbol{\mu}\|$ is inconsistent. Thus the "best attractor" that minimizes $\|\overline{\boldsymbol{x}}_{k,j} - \boldsymbol{\mu}\|$ for $k = 0$ is inconsistent by Lopuhaä (1999). See Theorem 3.20 a).

v) The classical estimator with breakdown $1/n$ is applied to each elemental start. Hence $\gamma_n \leq K/n \to 0$ as $n \to \infty$. □

Since the Fast-MCD estimator is a zero breakdown elemental concentration algorithm, the Hubert et al. (2008, 2012) claim that "MCD can be efficiently computed with the FAST-MCD estimator" is false. The Det-MCD estimator is a concentration algorithm using several intelligently selected starts. Fast-MCD and Det-MCD use iteration until convergence, and neither of these two estimators have been proven to be consistent or inconsistent. See Remark 3.4. The breakdown value of Det-MCD is also unknown.

Theorem 3.17. Neither Fast-MCD nor Det-MCD is the MCD estimator.

Proof. A necessary condition for an estimator to be the MCD estimator is that the determinant of the covariance matrix for the estimator be the smallest for every run in a simulation. Sometimes Fast-MCD had the smaller determinant and sometimes Det-MCD had the smaller determinant in the simulations done by Hubert et al. (2012). □

Remark 3.5. Let γ_o be the highest percentage of large outliers that an elemental concentration algorithm can detect reliably. For many data sets,

$$\gamma_o \approx \min\left(\frac{n-c_n}{n}, 1-[1-(0.2)^{1/K}]^{1/h}\right)100\% \tag{3.23}$$

if n is large, $c_n \geq n/2$ and $h = p+1$.

Proof. Suppose that the data set contains n cases with d outliers and $n-d$ clean cases. Suppose K elemental sets are chosen with replacement. If W_i is the number of outliers in the ith elemental set, then the W_i are iid hypergeometric$(d, n-d, h)$ random variables. Suppose that it is desired to find K such that the probability P(that at least one of the elemental sets is clean) $\equiv P_1 \approx 1-\alpha$ where $0 < \alpha < 1$. Then $P_1 = 1-$ P(none of the K elemental sets is clean) $\approx 1-[1-(1-\gamma)^h]^K$ by independence. If the contamination proportion γ is fixed, then the probability of obtaining at least one clean subset of size h with high probability (say $1-\alpha = 0.8$) is given by $0.8 = 1-[1-(1-\gamma)^h]^K$. Fix the number of starts K and solve this equation for γ. □

Equation (3.23) agrees very well with the Rousseeuw and Van Driessen (1999) simulation performed on the hybrid FMCD algorithm that uses both concentration and partitioning. Section 3.7 will provide theory for some useful practical algorithms.

3.7 Theory for Practical Estimators

This section presents the FCH, RFCH, and RMVN estimators. Recall from Definition 3.19 that a *concentration algorithm* uses K_n *starts* $(T_{-1,j}, \boldsymbol{C}_{-1,j})$. After finding $(T_{0,j}, \boldsymbol{C}_{0,j})$, each start is refined with k concentration steps, re-

sulting in K_n *attractors* $(T_{k,j}, \boldsymbol{C}_{k,j})$, and the concentration attractor $(T_A, \boldsymbol{C}_A)$ is the attractor that optimizes the criterion. Using $k = 10$ concentration steps works well.

The DGK estimator (Devlin et al. 1975, 1981) defined below is one example of a concentration algorithm estimator. The DGK estimator is affine equivariant since the classical estimator is affine equivariant and Mahalanobis distances are invariant under affine transformations by Theorem 3.11. This section will show that the Olive (2004a) MB estimator is a high breakdown estimator and that the DGK estimator is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$, the same quantity estimated by the MCD estimator. Both estimators use the classical estimator computed from $c_n \approx n/2$ cases. The breakdown point of the DGK estimator has been conjectured to be "at most $1/p$." See Rousseeuw and Leroy (1987, p. 254).

Definition 3.20. The *DGK estimator* $(T_{k,D}, \boldsymbol{C}_{k,D}) = (T_{DGK}, \boldsymbol{C}_{DGK})$ uses the classical estimator $(T_{-1,D}, \boldsymbol{C}_{-1,D}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ as the only start.

Definition 3.21. The *median ball (MB) estimator* $(T_{k,M}, \boldsymbol{C}_{k,M}) = (T_{MB}, \boldsymbol{C}_{MB})$ uses $(T_{-1,M}, \boldsymbol{C}_{-1,M}) = (\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ as the only start where $\text{MED}(\boldsymbol{W})$ is the coordinatewise median. So $(T_{0,M}, \boldsymbol{C}_{0,M})$ is the classical estimator applied to the "half set" of data closest to $\text{MED}(\boldsymbol{W})$ in Euclidean distance.

The proof of the following theorem implies that a high breakdown estimator $(T, \boldsymbol{C})$ has $\text{MED}(D_i^2) \leq V$ and that the hyperellipsoid $\{\boldsymbol{x} | D_{\boldsymbol{x}}^2 \leq D_{(c_n)}^2\}$ that contains $c_n \approx n/2$ of the cases is in some ball about the origin of radius r, where V and r do not depend on the outliers even if the number of outliers is close to $n/2$. Also the attractor of a high breakdown estimator is a high breakdown estimator if the number of concentration steps k is fixed, e.g. $k = 10$. The theorem implies that the MB estimator $(T_{MB}, \boldsymbol{C}_{MB})$ is high breakdown.

Theorem 3.18. Suppose $(T, \boldsymbol{C})$ is a high breakdown estimator where $\boldsymbol{C}$ is a symmetric, positive definite $p \times p$ matrix if the contamination proportion d_n/n is less than the breakdown value. Then the concentration attractor $(T_k, \boldsymbol{C}_k)$ is a high breakdown estimator if the coverage $c_n \approx n/2$ and the data are in general position.

Proof. Following Leon (1986, p. 280), if $\boldsymbol{A}$ is a symmetric positive definite matrix with eigenvalues $\tau_1 \geq \cdots \geq \tau_p$, then for any nonzero vector $\boldsymbol{x}$,

$$0 < \|\boldsymbol{x}\|^2 \, \tau_p \leq \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \leq \|\boldsymbol{x}\|^2 \, \tau_1. \tag{3.24}$$

Let $\lambda_1 \geq \cdots \geq \lambda_p$ be the eigenvalues of $\boldsymbol{C}$. By (3.24),

$$\frac{1}{\lambda_1}\|\boldsymbol{x} - T\|^2 \leq (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) \leq \frac{1}{\lambda_p}\|\boldsymbol{x} - T\|^2. \tag{3.25}$$

By (3.25), if the $D^2_{(i)}$ are the order statistics of the $D^2_i(T, \boldsymbol{C})$, then $D^2_{(i)} < V$ for some constant V that depends on the clean data but not on the outliers even if i and d_n are near $n/2$. (Note that $1/\lambda_p$ and $\text{MED}(\|\boldsymbol{x}_i - T\|^2)$ are both bounded for high breakdown estimators even for d_n near $n/2$.)

Following Johnson and Wichern (1988, pp. 50, 103), the boundary of the set $\{\boldsymbol{x}|D^2_{\boldsymbol{x}} \leq h^2\} = \{\boldsymbol{x}|(\boldsymbol{x} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x} - T) \leq h^2\}$ is a hyperellipsoid centered at T with axes of length $2h\sqrt{\lambda_i}$. Hence $\{\boldsymbol{x}|D^2_{\boldsymbol{x}} \leq D^2_{(c_n)}\}$ is contained in some ball about the origin of radius r where r does not depend on the number of outliers even for d_n near $n/2$. This is the set containing the cases used to compute $(T_0, \boldsymbol{C}_0)$. Since the set is bounded, T_0 is bounded and the largest eigenvalue $\lambda_{1,0}$ of $\boldsymbol{C}_0$ is bounded by Theorem 3.13. The determinant $det(\boldsymbol{C}_{MCD})$ of the HB minimum covariance determinant estimator satisfies $0 < det(\boldsymbol{C}_{MCD}) \leq det(\boldsymbol{C}_0) = \lambda_{1,0} \cdots \lambda_{p,0}$, and $\lambda_{p,0} > \inf det(\boldsymbol{C}_{MCD})/\lambda^{p-1}_{1,0} > 0$ where the infimum is over all possible data sets with $n - d_n$ clean cases and d_n outliers. Since these bounds do not depend on the outliers even for d_n near $n/2$, $(T_0, \boldsymbol{C}_0)$ is a high breakdown estimator. Now repeat the argument with $(T_0, \boldsymbol{C}_0)$ in place of $(T, \boldsymbol{C})$ and $(T_1, \boldsymbol{C}_1)$ in place of $(T_0, \boldsymbol{C}_0)$. Then $(T_1, \boldsymbol{C}_1)$ is high breakdown. Repeating the argument iteratively shows $(T_k, \boldsymbol{C}_k)$ is high breakdown. □

The following corollary shows that it is easy to find a subset J of $c_n \approx n/2$ cases such that the classical estimator $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ applied to J is a HB estimator of MLD. Note that $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J) = (T_0, \boldsymbol{C}_0)$ in the MB concentration algorithm.

Theorem 3.19. Let J consist of the c_n cases $\boldsymbol{x}_i$ such that $\|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\| \leq \text{MED}(\|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\|)$. Then the classical estimator $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ applied to J is a HB estimator of MLD.

To investigate the consistency and rate of robust estimators of multivariate location and dispersion, review Definitions 11.14 and 11.15.

The following assumption (E1) gives a class of distributions where we can prove that the new robust estimators are $\sqrt{n}$ consistent. Cator and Lopuhaä (2010, 2012) showed that MCD is consistent provided that the MCD functional is unique. Distributions where the functional is unique are called "unimodal," and rule out, for example, a spherically symmetric uniform distribution. Theorem 3.20 is crucial for theory and Theorem 3.21 shows that under (E1), both MCD and DGK are estimating $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$.

Assumption (E1): The $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid from a "unimodal" $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with nonsingular covariance matrix $\text{Cov}(\boldsymbol{x}_i)$ where g is continuously differentiable with finite 4th moment: $\int (\boldsymbol{x}^T\boldsymbol{x})^2 g(\boldsymbol{x}^T\boldsymbol{x})d\boldsymbol{x} < \infty$.

Lopuhaä (1999) showed that if a start $(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$, then the classical estimator applied to the cases with $D^2_i(T, \boldsymbol{C}) \leq h^2$ is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ where $a, s > 0$ are some constants. Affine equivariance is not used for $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. Also, the attrac-

tor and the start have the same rate. If the start is inconsistent, then so is the attractor. The weight function $I(D_i^2(T, \boldsymbol{C}) \leq h^2)$ is an indicator that is 1 if $D_i^2(T, \boldsymbol{C}) \leq h^2$ and 0 otherwise.

Theorem 3.20, Lopuhaä (1999). Assume the number of concentration steps k is fixed. a) If the start $(T, \boldsymbol{C})$ is inconsistent, then so is the attractor.

b) Suppose $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\boldsymbol{I}_p)$ with rate n^δ where $s > 0$ and $0 < \delta \leq 0.5$. Assume (E1) holds and $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. Then the classical estimator $(T_0, \boldsymbol{C}_0)$ applied to the cases with $D_i^2(T, \boldsymbol{C}) \leq h^2$ is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{I}_p)$ with the same rate n^δ where $a > 0$.

c) Suppose $(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^δ where $s > 0$ and $0 < \delta \leq 0.5$. Assume (E1) holds. Then the classical estimator $(T_0, \boldsymbol{C}_0)$ applied to the cases with $D_i^2(T, \boldsymbol{C}) \leq h^2$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with the same rate n^δ where $a > 0$. The constant a depends on the positive constants s, h, p, and the elliptically contoured distribution, but does not otherwise depend on the consistent start $(T, \boldsymbol{C})$.

Let $\delta = 0.5$. Applying Theorem 3.20c) iteratively for a fixed number k of steps produces a sequence of estimators $(T_0, \boldsymbol{C}_0), ..., (T_k, \boldsymbol{C}_k)$ where $(T_j, \boldsymbol{C}_j)$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_j\boldsymbol{\Sigma})$ where the constants $a_j > 0$ depend on s, h, p, and the elliptically contoured distribution, but do not otherwise depend on the consistent start $(T, \boldsymbol{C}) \equiv (T_{-1}, \boldsymbol{C}_{-1})$.

The 4th moment assumption was used to simplify theory, but likely holds under 2nd moments. Affine equivariance is needed so that the attractor is affine equivariant, but probably is not needed to prove consistency.

Conjecture 3.1. Change the finite 4th moments assumption to a finite 2nd moments in assumption E1). Suppose $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^δ where $s > 0$ and $0 < \delta \leq 0.5$. Then the classical estimator applied to the cases with $D_i^2(T, \boldsymbol{C}) \leq h^2$ is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with the same rate n^δ where $a > 0$.

Remark 3.6. To see that the Lopuhaä (1999) theory extends to concentration where the weight function uses $h^2 = D_{(c_n)}^2(T, \boldsymbol{C})$, note that $(T, \tilde{\boldsymbol{C}}) \equiv (T, D_{(c_n)}^2(T, \boldsymbol{C})\ \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, b\boldsymbol{\Sigma})$ where $b > 0$ is derived in (3.27), and weight function $I(D_i^2(T, \tilde{\boldsymbol{C}}) \leq 1)$ is equivalent to the concentration weight function $I(D_i^2(T, \boldsymbol{C}) \leq D_{(c_n)}^2(T, \boldsymbol{C}))$. As noted above Theorem 3.11, $(T, \tilde{\boldsymbol{C}})$ is affine equivariant if $(T, \boldsymbol{C})$ is affine equivariant. Hence Lopuhaä (1999) theory applied to $(T, \tilde{\boldsymbol{C}})$ with $h = 1$ is equivalent to theory applied to affine equivariant $(T, \boldsymbol{C})$ with $h^2 = D_{(c_n)}^2(T, \boldsymbol{C})$.

If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ with rate n^δ where $0 < \delta \leq 0.5$, then $D^2(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) =$

$$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)^T [\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1} + s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)$$

$$= s^{-1}D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(n^{-\delta}). \tag{3.26}$$

Thus the sample percentiles of $D_i^2(T, \boldsymbol{C})$ are consistent estimators of the percentiles of $s^{-1}D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Suppose $c_n/n \to \xi \in (0, 1)$ as $n \to \infty$, and let $D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the 100ξth percentile of the population squared distances. Then $D_{(c_n)}^2(T, \boldsymbol{C}) \xrightarrow{P} s^{-1}D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $b\boldsymbol{\Sigma} = s^{-1}D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})s\boldsymbol{\Sigma} = D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})\boldsymbol{\Sigma}$. Thus

$$b = D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{3.27}$$

does not depend on $s > 0$ or $\delta \in (0, 0.5]$. □

Concentration applies the classical estimator to cases with $D_i^2(T, \boldsymbol{C}) \leq D_{(c_n)}^2(T, \boldsymbol{C})$. Let $c_n \approx n/2$ and

$$b = D_{0.5}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

be the population median of the population squared distances. By Remark 3.6, if $(T, \boldsymbol{C})$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ then $(T, \tilde{\boldsymbol{C}}) \equiv (T, D_{(c_n)}^2(T, \boldsymbol{C})\ \boldsymbol{C})$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, b\boldsymbol{\Sigma})$, and $D_i^2(T, \tilde{\boldsymbol{C}}) \leq 1$ is equivalent to $D_i^2(T, \boldsymbol{C}) \leq D_{(c_n)}^2(T, \boldsymbol{C})$). Hence Lopuhaä (1999) theory applied to $(T, \tilde{\boldsymbol{C}})$ with $h = 1$ is equivalent to theory applied to the concentration estimator using the affine equivariant estimator $(T, \boldsymbol{C}) \equiv (T_{-1}, \boldsymbol{C}_{-1})$ as the start. Since b does not depend on s, concentration produces a sequence of estimators $(T_0, \boldsymbol{C}_0), ..., (T_k, \boldsymbol{C}_k)$ where $(T_j, \boldsymbol{C}_j)$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ where the constant $a > 0$ is the same for $j = 0, 1, ..., k$.

Theorem 3.21 shows that $a = a_{MCD}$ where $\xi = 0.5$. Hence concentration with a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^δ as a start results in a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ with rate n^δ. This result can be applied iteratively for a finite number of concentration steps. Hence DGK is a $\sqrt{n}$ consistent affine equivariant estimator of the same quantity that MCD is estimating. It is not known if the results hold if concentration is iterated to convergence. For multivariate normal data, $D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \chi_p^2$.

Theorem 3.21. Assume that (E1) holds and that $(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^δ where the constants $s > 0$ and $0 < \delta \leq 0.5$. Then the classical estimator $(\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$ computed from the $c_n \approx n/2$ of cases with the smallest distances $D_i(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ with the same rate n^δ.

Proof. By Remark 3.6 the estimator is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with rate n^δ. By the remarks above, a will be the same for any consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ and a does not depend on $s > 0$ or $\delta \in (0, 0.5]$. Hence the result follows if $a = a_{MCD}$. The MCD estimator is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ by Cator and Lopuhaä (2010, 2012). If the MCD estimator is the start, then it

is also the attractor by Rousseeuw and Van Driessen (1999) who show that concentration does not increase the MCD criterion. Hence $a = a_{MCD}$. □

Next we define the easily computed robust $\sqrt{n}$ consistent FCH estimator, so named since it is fast, consistent, and uses a high breakdown attractor. The FCH and MBA estimators use the $\sqrt{n}$ consistent DGK estimator $(T_{DGK}, \boldsymbol{C}_{DGK})$ and the high breakdown MB estimator $(T_{MB}, \boldsymbol{C}_{MB})$ as attractors.

Definition 3.22. Let the "median ball" be the hypersphere containing the "half set" of data closest to MED($\boldsymbol{W}$) in Euclidean distance. The *FCH estimator* uses the MB attractor if the DGK location estimator T_{DGK} is outside of the median ball, and the attractor with the smallest determinant, otherwise. Let $(T_A, \boldsymbol{C}_A)$ be the attractor used. Then the estimator $(T_{FCH}, \boldsymbol{C}_{FCH})$ takes $T_{FCH} = T_A$ and

$$\boldsymbol{C}_{FCH} = \frac{\text{MED}(D_i^2(T_A, \boldsymbol{C}_A))}{\chi^2_{p,0.5}} \boldsymbol{C}_A \tag{3.28}$$

where $\chi^2_{p,0.5}$ is the 50th percentile of a chi–square distribution with p degrees of freedom.

Remark 3.7. The *MBA estimator* $(T_{MBA}, \boldsymbol{C}_{MBA})$ uses the attractor $(T_A, \boldsymbol{C}_A)$ with the smallest determinant. Hence the DGK estimator is used as the attractor if $det(\boldsymbol{C}_{DGK}) \leq det(\boldsymbol{C}_{MB})$, and the MB estimator is used as the attractor, otherwise. Then $T_{MBA} = T_A$ and $\boldsymbol{C}_{MBA}$ is computed using the right hand side of (3.28). The difference between the FCH and MBA estimators is that the FCH estimator also uses a location criterion to choose the attractor: if the DGK location estimator T_{DGK} has a greater Euclidean distance from MED($\boldsymbol{W}$) than half the data, then FCH uses the MB attractor. The FCH estimator only uses the attractor with the smallest determinant if $\|T_{DGK} - \text{MED}(\boldsymbol{W})\| \leq \text{MED}(D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p))$. Using the location criterion increases the outlier resistance of the FCH estimator for certain types of outliers, as will be seen in Section 3.9.

The following theorem shows the FCH estimator has good statistical properties. We conjecture that FCH is high breakdown. Note that the location estimator T_{FCH} is high breakdown and that $det(\boldsymbol{C}_{FCH})$ is bounded away from 0 and ∞ if the data is in general position, even if nearly half of the cases are outliers.

Theorem 3.22. T_{FCH} is high breakdown if the clean data are in general position. Suppose (E1) holds. If $(T_A, \boldsymbol{C}_A)$ is the DGK or MB attractor with the smallest determinant, then $(T_A, \boldsymbol{C}_A)$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$. Hence the MBA and FCH estimators are outlier resistant $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ where $c = u_{0.5}/\chi^2_{p,0.5}$, and $c = 1$ for multivariate normal data.

Proof. T_{FCH} is high breakdown since it is a bounded distance from MED($\boldsymbol{W}$) even if the number of outliers is close to $n/2$. Under (E1) the FCH and MBA estimators are asymptotically equivalent since $\|T_{DGK} - \text{MED}(\boldsymbol{W})\| \to 0$ in probability. The estimator satisfies $0 < det(\boldsymbol{C}_{MCD}) \leq det(\boldsymbol{C}_A) \leq det(\boldsymbol{C}_{0,M}) < \infty$ by Theorem 3.18 if up to nearly 50% of the cases are outliers. If the distribution is spherical about $\boldsymbol{\mu}$, then the result follows from Pratt (1959) and Theorem 3.14 since both starts are $\sqrt{n}$ consistent. Otherwise, the MB estimator $\boldsymbol{C}_{MB}$ is a biased estimator of $a_{MCD}\boldsymbol{\Sigma}$. But the DGK estimator $\boldsymbol{C}_{DGK}$ is a $\sqrt{n}$ consistent estimator of $a_{MCD}\boldsymbol{\Sigma}$ by Theorem 3.21 and $\|\boldsymbol{C}_{MCD} - \boldsymbol{C}_{DGK}\| = O_P(n^{-1/2})$. Thus the probability that the DGK attractor minimizes the determinant goes to one as $n \to \infty$, and $(T_A, \boldsymbol{C}_A)$ is asymptotically equivalent to the DGK estimator $(T_{DGK}, \boldsymbol{C}_{DGK})$.

Let $\boldsymbol{C}_F = \boldsymbol{C}_{FCH}$ or $\boldsymbol{C}_F = \boldsymbol{C}_{MBA}$. Let $P(U \leq u_\alpha) = \alpha$ where U is given by (3.9). Then the scaling in (3.28) makes $\boldsymbol{C}_F$ a consistent estimator of $c\boldsymbol{\Sigma}$ where $c = u_{0.5}/\chi^2_{p,0.5}$, and $c = 1$ for multivariate normal data. □

A standard method of reweighting can be used to produce the RMBA and RFCH estimators. RMVN uses a slightly modified method of reweighting so that RMVN gives good estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for multivariate normal data, even when certain types of outliers are present.

Definition 3.23. The *RFCH estimator* uses two standard reweighting steps. Let $(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1)$ be the classical estimator applied to the n_1 cases with $D_i^2(T_{FCH}, \boldsymbol{C}_{FCH}) \leq \chi^2_{p,0.975}$, and let

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{\text{MED}(D_i^2(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1))}{\chi^2_{p,0.5}} \tilde{\boldsymbol{\Sigma}}_1.$$

Then let $(T_{RFCH}, \tilde{\boldsymbol{\Sigma}}_2)$ be the classical estimator applied to the cases with $D_i^2(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1) \leq \chi^2_{p,0.975}$, and let

$$\boldsymbol{C}_{RFCH} = \frac{\text{MED}(D_i^2(T_{RFCH}, \tilde{\boldsymbol{\Sigma}}_2))}{\chi^2_{p,0.5}} \tilde{\boldsymbol{\Sigma}}_2.$$

RMBA and RFCH are $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ by Lopuhaä (1999) where the weight function uses $h^2 = \chi^2_{p,0.975}$, but the two estimators use nearly 97.5% of the cases if the data is multivariate normal.

Definition 3.24. The *RMVN estimator* uses $(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1)$ and n_1 as above. Let $q_1 = \min\{0.5(0.975)n/n_1, 0.995\}$, and

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{\text{MED}(D_i^2(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1))}{\chi^2_{p,q_1}} \tilde{\boldsymbol{\Sigma}}_1.$$

Then let $(T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2)$ be the classical estimator applied to the n_2 cases with $D_i^2(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1)) \leq \chi^2_{p,0.975}$. Let $q_2 = \min\{0.5(0.975)n/n_2, 0.995\}$, and

$$\boldsymbol{C}_{RMVN} = \frac{\text{MED}(D_i^2(T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2))}{\chi^2_{p,q_2}} \tilde{\boldsymbol{\Sigma}}_2.$$

The RMVN estimator is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$ by Lopuhaä (1999) where the weight function uses $h^2 = \chi^2_{p,0.975}$ and $d = u_{0.5}/\chi^2_{p,q}$ where $q_2 \to q$ in probability as $n \to \infty$. Here $0.5 \leq q < 1$ depends on the elliptically contoured distribution, but $q = 0.5$ and $d = 1$ for multivariate normal data.

If the bulk of the data is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the RMVN estimator can give useful estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for certain types of outliers where FCH and RFCH estimate $(\boldsymbol{\mu}, d_E\boldsymbol{\Sigma})$ for $d_E > 1$. To see this claim, let $0 \leq \gamma < 0.5$ be the outlier proportion. If $\gamma = 0$, then $n_i/n \xrightarrow{P} 0.975$ and $q_i \xrightarrow{P} 0.5$. If $\gamma > 0$, suppose the outlier configuration is such that the $D_i^2(T_{FCH}, \boldsymbol{C}_{FCH})$ are roughly χ^2_p for the clean cases, and the outliers have larger D_i^2 than the clean cases. Then $\text{MED}(D_i^2) \approx \chi^2_{p,q}$ where $q = 0.5/(1-\gamma)$. For example, if $n = 100$ and $\gamma = 0.4$, then there are 60 clean cases, $q = 5/6$, and the quantile $\chi^2_{p,q}$ is being estimated instead of $\chi^2_{p,0.5}$. Now $n_i \approx n(1-\gamma)0.975$, and q_i estimates q. Thus $\boldsymbol{C}_{RMVN} \approx \boldsymbol{\Sigma}$. Of course consistency cannot generally be claimed when outliers are present.

Remark 3.8. The FCH, RFCH, and RMVN estimators may be the only practical MLD estimators that have been shown to be $\sqrt{n}$ consistent on a large class of distributions and highly outlier resistant. The MBA and RMBA estimators have also been shown to be $\sqrt{n}$ consistent, but have less outlier resistance. The **main competitors** for the Olive and Hawkins (2010) FCH, RFCH, and RMVN estimators are the Maronna and Zamar (2002) *OGK estimator*, the Hubert et al. (2012) *Det-MCD estimator* which have not been proven to be consistent or positive breakdown, and the *Sign Covariance Matrix* shown to be high breakdown by Croux et al. (2010). Also see Taskinen et al. (2012). Croux et al. (2010) showed that the practical Sign Covariance Matrix and k-step Spatial Sign Covariance Matrix are high breakdown. They claimed that under regularity conditions, these two estimators consistently estimate the orientation of the dispersion matrix.

Estimators with complexity higher than $O[(n^3+n^2p+np^2+p^3)\log(n)]$ take too long to compute and will rarely be used. Reyen et al. (2009) simulated the OGK and the Olive (2004a) median ball algorithm (MBA) estimators for $p = 100$ and n up to 50000, and noted that the OGK complexity is $O[p^3+np^2\log(n)]$ while that of MBA is $O[p^3+np^2+np\log(n)]$. FCH, RMBA, and RMVN have the same complexity as MBA. Fast-MCD has the same complexity as FCH, but FCH is roughly 100 to 200 times faster.

The fastest estimators of multivariate location and dispersion that have been shown to be both consistent and high breakdown are the MCD estimator with $O(n^v)$ complexity where $v = 1+p(p+3)/2$ and possibly an all elemental

subset estimator of He and Wang (1997). See Bernholt and Fischer (2004). The minimum volume ellipsoid estimator complexity is far higher, and **for $p > 2$ there may be no known method for computing** S, τ, projection based, and constrained M estimators. For some depth estimators, like the Stahel-Donoho estimator, the exact algorithm of Liu and Zuo (2014) appears to take too long if $p \geq 6$ and $n \geq 100$, and simulations may need $p \leq 3$. □

Remark 3.9. Practical consistent highly outlier resistant estimators are still affected by certain types of outliers. The median ball and location criterion give FCH, RFCH, and RMVN considerable outlier resistance to outlier configurations that lie outside the "median ball," including outlier configurations that can cause problems for the MCD estimator. For p not much larger than 5, the elemental concentration algorithm with the MCD criterion can detect some outlier types that are not detected by FCH, RFCH, and RMVN. These outlier types tend to be within the "median ball." The point mass outlier configuration, where all of the outliers are equal to $\boldsymbol{x}_O$, often causes numerical problems. The OGK and MB estimators have considerable resistance to point mass outliers. The DGK, Fast-MCD, Det-MCD, and MCD estimators have problems with the point mass. Suppose the bulk of the data lies in a hyperellipsoid. A 40% point mass can combine with 10% of the clean data to form a hyperellipsoid covering half of the data with smaller volume than a hyperellipsoid covering half of the data without any outliers. Then the MCD criterion tends to select a "half set" that contains the outliers. The location criterion used by the FCH estimator will often reject the DGK attractor for the point mass. However, the current program for FCH fails if the DGK estimator can't be computed, which often happens for the point mass. For a single data set, just use the scaled MB estimator if the DGK estimator causes the FCH, RFCH, or RMVN program to fail. It would be nice to have a program that that does not fail when the DGK estimator fails. Since the point mass causes numerical difficulties for most estimators, simulations often use a near point mass: the outliers are tightly clustered about a single point $\boldsymbol{x}_O$, but the outliers have a nonsingular covariance matrix.

Table 3.1 Average Dispersion Matrices for Near Point Mass Outliers

RMVN	FMCD	OGK	MB
$\begin{bmatrix} 1.002 & -0.014 \\ -0.014 & 2.024 \end{bmatrix}$	$\begin{bmatrix} 0.055 & 0.685 \\ 0.685 & 122.5 \end{bmatrix}$	$\begin{bmatrix} 0.185 & 0.089 \\ 0.089 & 36.24 \end{bmatrix}$	$\begin{bmatrix} 2.570 & -0.082 \\ -0.082 & 5.241 \end{bmatrix}$

Table 3.2 Average Dispersion Matrices for Mean Shift Outliers

RMVN	FMCD	OGK	MB
$\begin{bmatrix} 0.990 & 0.004 \\ 0.004 & 2.014 \end{bmatrix}$	$\begin{bmatrix} 2.530 & 0.003 \\ 0.003 & 5.146 \end{bmatrix}$	$\begin{bmatrix} 19.67 & 12.88 \\ 12.88 & 39.72 \end{bmatrix}$	$\begin{bmatrix} 2.552 & 0.003 \\ 0.003 & 5.118 \end{bmatrix}$

Simulations suggested $(T_{RMVN}, \boldsymbol{C}_{RMVN})$ gives useful estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for a variety of outlier configurations. Using 20 runs and $n = 1000$, the averages of the dispersion matrices were computed when the bulk of the data are iid $N_2(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = diag(1, 2)$. For clean data, FCH, RFCH, and RMVN give $\sqrt{n}$ consistent estimators of $\boldsymbol{\Sigma}$, while Fast-MCD (FMCD) and the OGK estimator seem to be approximately unbiased for $\boldsymbol{\Sigma}$. The median ball estimator was scaled using (3.28) and estimated $diag(1.13, 1.85)$.

Next the data had $\gamma = 0.4$ and the outliers had $\boldsymbol{x} \sim N_2((0, 15)^T, 0.0001\boldsymbol{I}_2)$, a near point mass at the major axis. FCH, MB, and RFCH estimated $2.6\boldsymbol{\Sigma}$ while RMVN estimated $\boldsymbol{\Sigma}$. FMCD and OGK failed to estimate d $\boldsymbol{\Sigma}$. Note that $\chi^2_{2,5/6}/\chi^2_{2,0.5} = 2.585$. See Table 3.1. The following R commands were used where `mldsim` is from *rpack*.

```
qchisq(5/6,2)/qchisq(.5,2) = 2.584963
mldsim(n=1000,p=2,outliers=6,pm=15)
```

Next the data had $\gamma = 0.4$ and the outliers had $\boldsymbol{x} \sim N_2((20, 20)^T, \boldsymbol{\Sigma})$, a mean shift with the same covariance matrix as the clean cases. Rocke and Woodruff (1996) suggest that outliers with mean shift are hard to detect. FCH, FMCD, MB, and RFCH estimated $2.6\boldsymbol{\Sigma}$ while RMVN estimated $\boldsymbol{\Sigma}$, and OGK failed. See Table 3.2. The *R command* is shown below.

```
mldsim(n=1000,p=2,outliers=3,pm=20)
```

Remark 3.10. The RFCH and RMVN estimators are recommended. If these estimators are too slow and outlier detection is of interest, try the RMB estimator, the reweighted MB estimator. If RMB is too slow or if $n < 2(p+1)$, the Euclidean distances $D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I})$ of $\boldsymbol{x}_i$ from the coordinatewise median MED$(\boldsymbol{W})$ may be useful. A DD plot of $D_i(\overline{\boldsymbol{x}}, \boldsymbol{I})$ versus $D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I})$ is also useful for outlier detection and for whether $\overline{\boldsymbol{x}}$ and MED$(\boldsymbol{W})$ are giving similar estimates of multivariate location. See Section 3.10. For DD plots, see Section 3.8.

Example 3.4. Tremearne (1911) recorded *height* = x[,1] and *height while kneeling* = x[,2] of 112 people. Figure 3.2a shows a scatterplot of the data. Case 3 has the largest Euclidean distance of 214.767 from MED$(\boldsymbol{W})$ $= (1680, 1240)^T$, but if the distances correspond to the contours of a covering ellipsoid, then case 44 has the largest distance. For $k = 0$, $(T_{0,M}, \boldsymbol{C}_{0,M})$ is the classical estimator applied to the "half set" of cases closest to MED$(\boldsymbol{W})$ in Euclidean distance. The hypersphere (circle) centered at MED$(\boldsymbol{W})$ that covers half the data is small because the data density is high near MED$(\boldsymbol{W})$. The median Euclidean distance is 59.661 and case 44 has Euclidean distance 77.987. Hence the intersection of the sphere and the data is a highly correlated clean ellipsoidal region. Figure 3.2b shows the DD plot of the classical distances versus the MB distances. Notice that both the classical and MB estimators give the largest distances to cases 3 and 44. Notice that case 44 could not be detected using marginal methods.

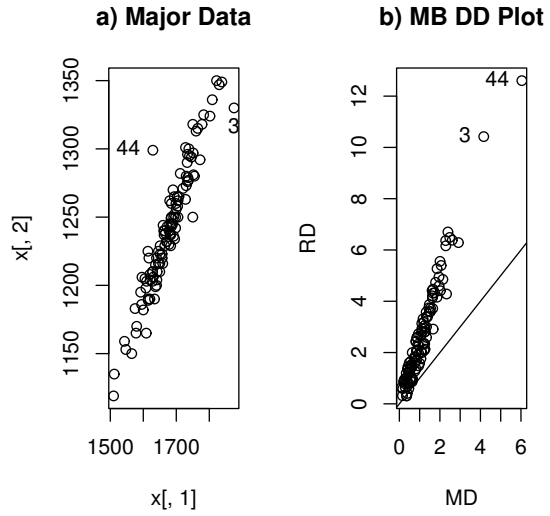

Fig. 3.2 Plots for Major Data

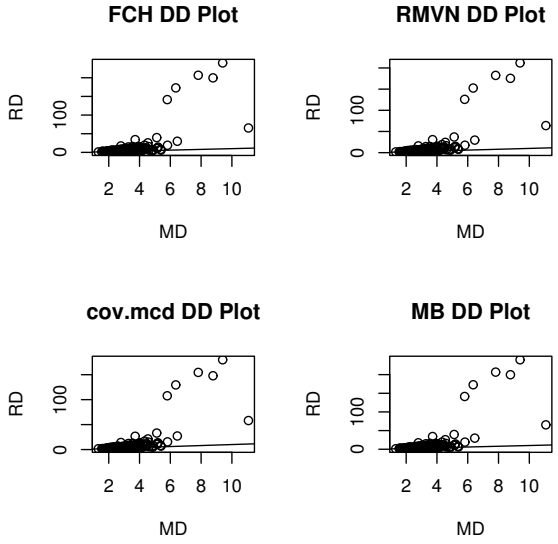

Fig. 3.3 DD Plots for Gladstone Data

As the dimension p gets larger, outliers that can not be detected by marginal methods (case 44 in Example 3.4) become harder to detect. When $p = 3$ imagine that the clean data is a baseball bat or stick with one end at the SW corner of the bottom of the box (corresponding to the coordinate axes) and one end at the NE corner of the top of the box. If the outliers are a ball, there is much more room to hide them in the box than in a covering rectangle when $p = 2$.

Example 3.5. The estimators can be useful when the data is not elliptically contoured. The Gladstone (1905) data has 11 variables on 267 persons after death. Head measurements were *breadth, circumference, head height, length,* and *size* as well as *cephalic index* and *brain weight*. *Age*, *height*, and two categorical variables *ageclass* (0: under 20, 1: 20-45, 2: over 45) and *sex* were also given. Figure 3.3 shows the DD plots for the FCH, RMVN, `cov.mcd`, and MB estimators. The DD plots from the DGK, MBA, and RFCH estimators were similar, and the six outliers in Figure 3.3 correspond to the six infants in the data set.

3.8 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)

The classical Mahalanobis distance will be denoted by MD_i, and corresponds to the sample mean and sample covariance matrix $(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ of Definition 3.10. When $T(\boldsymbol{W})$ and $\boldsymbol{C}(\boldsymbol{W})$ are estimators other than the sample mean and covariance, $D_i = \sqrt{D_i^2}$ will sometimes be denoted by RD_i.

Definition 3.25: 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 theorem 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 100αth sample percentile of the MD_i. Nevertheless, the variability in the DD plot may increase with the distances. Let $K > 0$ be a constant, e.g. the 99th percentile of the χ^2_p distribution.

Theorem 3.23. 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): $D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) =$

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

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

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

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

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

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

$$+\frac{1}{a_j}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_j)^T[a_j\hat{\boldsymbol{\Sigma}}_j^{-1}-\boldsymbol{\Sigma}^{-1}](\boldsymbol{x}-\hat{\boldsymbol{\mu}}_j) \tag{3.29}$$

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. □

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 Theorem 3.23, 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 Theorem 3.23, 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 3.22 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 et al. (1997), and Liu et al. (1999) for references.

Example 3.6. Rousseeuw and Van Driessen (1999) chose 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

$$\mathrm{RD}_i = \frac{\sqrt{\chi^2_{p,0.5}}}{\mathrm{med}(D_i(A))} D_i(A) \quad \text{so that} \quad \tau = \frac{\sqrt{\chi^2_{\mathrm{p},0.5}}}{\mathrm{med}(\mathrm{D_i(A)})}. \tag{3.30}$$

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 Equation (3.10)), then use

$$\mathrm{RD}_i = \tau \; D_i(A) = \frac{\mathrm{MED}}{\mathrm{med}(D_i(A))} D_i(A). \tag{3.31}$$

We recommend using RFCH or RMVN as the robust estimators in DD plots. 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 hyperellipsoids that are nearly concentric. We suspect that the situation worsens as p increases. The `cov.mcd` estimator is basically an implementation of the elemental FMCD concentration algorithm described in Section 3.6. 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 3.2. 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 under mild regularity conditions 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

$$\left[\frac{\mathrm{med}(D_i(A))}{\mathrm{med}(\mathrm{MD}_i)}\right]^2 \boldsymbol{C}_A \tag{3.32}$$

estimates $\mathrm{Cov}(\boldsymbol{x})$. 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 Theorem 3.23 if consistent estimators are used.

In a simulation study, $N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ data were generated and `cov.mcd` was used to compute first the $D_i(A)$, and then the RD_i using Equation (3.30). The results are shown in Table 3.3. 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

Table 3.3 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

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 nonsingular estimator of multivariate location and dispersion defines a hyperellipsoid, the DD plot can be used to examine which points are in the robust hyperellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - T_R)^T \boldsymbol{C}_R^{-1} (\boldsymbol{x} - T_R) \leq RD^2_{(h)}\} \tag{3.33}$$

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

$$\{\boldsymbol{x} : (\boldsymbol{x} - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x} - \overline{\boldsymbol{x}}) \leq MD^2_{(h)}\}. \tag{3.34}$$

In the DD plot, points below $RD_{(h)}$ correspond to cases that are in the hyperellipsoid given by Equation (3.33) while points to the left of $MD_{(h)}$ are in a hyperellipsoid determined by Equation (3.34). Hence the DD plot can be used to visualize the prediction regions of Section 5.1.

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

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

for $i = 1, ..., n$. When the distribution is not EC, the RMVN (or RFCH or FMCD) estimator and $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ will often produce hyperellipsoids that are far from concentric.

Application 3.1. 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.

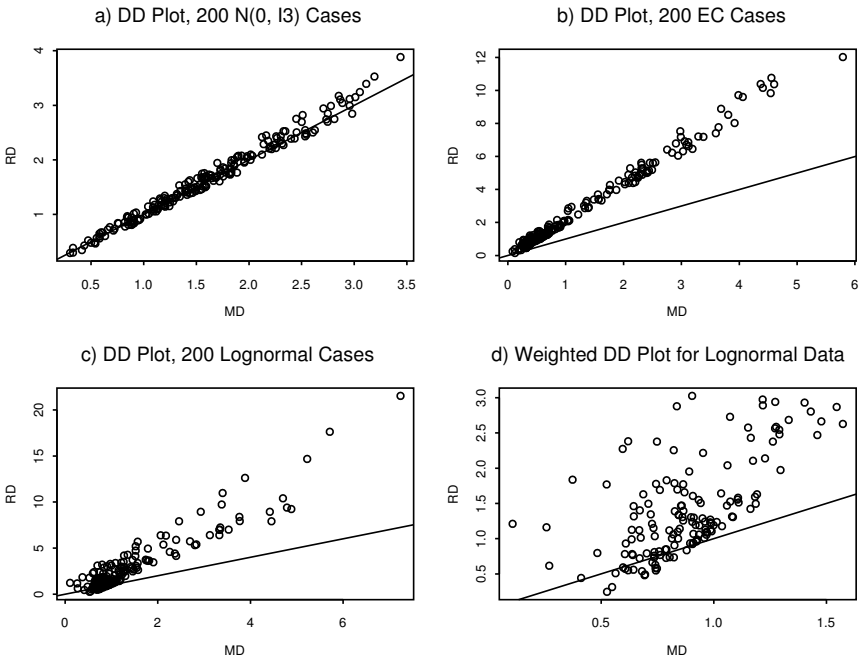

Fig. 3.4 4 DD Plots

For this application, the RFCH and RMVN estimators 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 3.4 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 3.4a 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 3.4b 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 $\boldsymbol{\Sigma}$, Theorem 3.23 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 3.4b are highly correlated and still follow a line through the origin with a slope close to 2.0.

Figures 3.4c and 3.4d illustrate how to use the weighted DD plot. The ith case in Figure 3.4c is $(\exp(x_{i,1}), \exp(x_{i,2}), \exp(x_{i,3}))^T$ where $\boldsymbol{x}_i$ is the ith case in Figure 3.4a; i.e. 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 3.4d 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 3.4d 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) gave 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 3.7. Buxton (1920, pp. 232-5) gave 20 measurements of 88 men. We will examine whether the multivariate normal distribution is a reasonable model for the measurements *head length, nasal height, bigonal breadth,* and *cephalic index* where one case has been deleted due to missing values. Figure 3.5a shows the DD plot. Five head lengths were recorded to be around 5 feet and are massive outliers. Figure 3.5b is the DD plot computed after

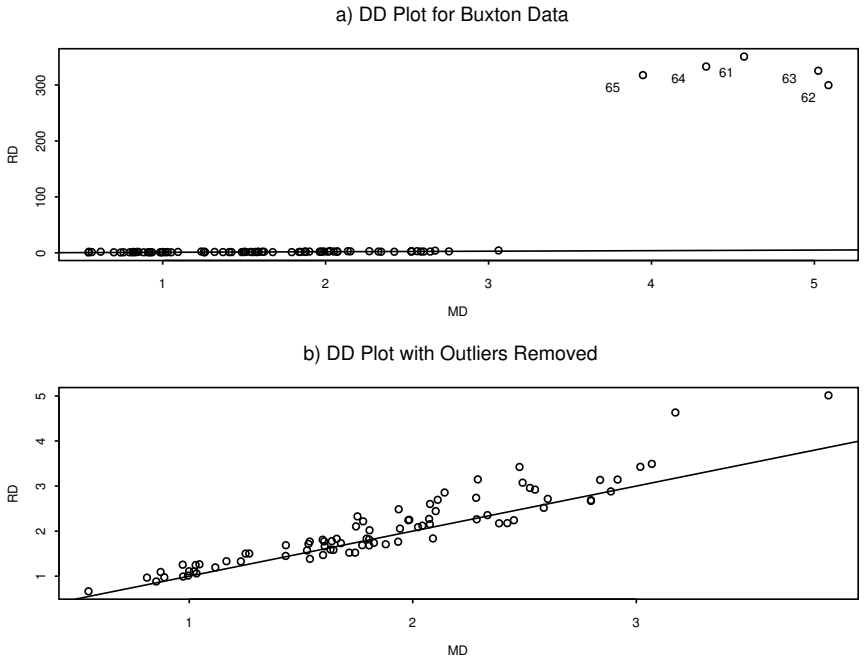

Fig. 3.5 DD Plots for the Buxton Data

deleting these points and suggests that the multivariate normal distribution is reasonable. (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 3.4a, then perhaps no transformation is needed. For the data in Figure 3.4c, a good numerical procedure should suggest coordinatewise log transforms. Following this transformation, the resulting plot shown in Figure 3.4a indicates that the transformation to normality was successful.

Application 3.2. The DD plot can be used to detect multivariate outliers. See Figures 3.2, 3.3, 3.5a, and 3.6.

Warning: It is important to know that plots fill space. If there is a single outlier, then often it will appear in the upper left or upper right corner of the DD plot, where RD is large, since the plot has to cover the outlier. The rest of the data will often appear to be tightly clustered about the identity

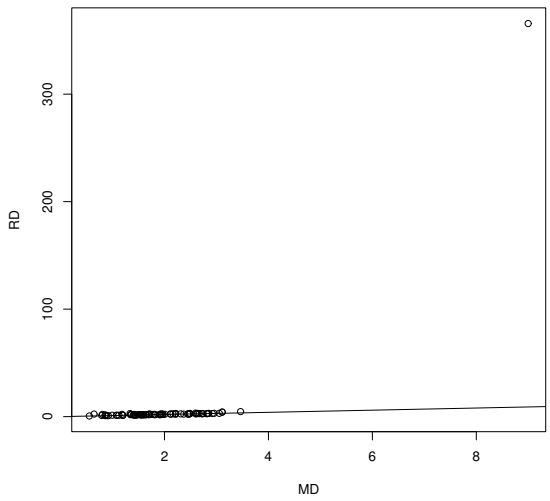

Fig. 3.6 DD Plot With One Outlier in the Upper Right Corner

line. Beginners sometimes fail to spot the single outlier because they do not know that the plot will fill space. There is a lot of blank space because of the outlier. If the outlier was not present, then the box would not extend much above the identity line in the upper right corner of the plot. For example, suppose all of the outliers except point 63 were deleted from the Buxton data. Then compare the DD plot in Figure 3.5 b) where all of the outliers have been deleted, with the DD plot in Figure 3.6 where the single outlier is in the upper right corner. *R* commands to produce Figures 3.5 and 3.6 are shown below.

```
library(MASS)
x <- cbind(buxy,buxx)
ddplot(x,type=3) #Figure 3.5a), right click Stop

zx <- x[-c(61:65),]
ddplot(zx,type=3) #Figure 3.5b), right click Stop

zz <- x[-c(61,62,64,65),]
ddplot(zz,type=3) #Figure 3.6, right click Stop
```

3.9 Outlier Resistance and Simulations

```
RMVN                                FMCD
 0.996  0.014  0.002 -0.001   0.931  0.017  0.011 0.000
 0.014  2.012 -0.001  0.029   0.017  1.885 -0.003 0.022
 0.002 -0.001  2.984  0.003   0.011 -0.003  2.803 0.010
-0.001  0.029  0.003  3.994   0.000  0.022  0.010 3.752
```

Simulations were used to compare $(T_{FCH}, \boldsymbol{C}_{FCH})$, $(T_{RFCH}, \boldsymbol{C}_{RFCH})$, $(T_{RMVN}, \boldsymbol{C}_{RMVN})$, and $(T_{FMCD}, \boldsymbol{C}_{FMCD})$. Shown above are the averages, using 20 runs and $n = 1000$, of the dispersion matrices when the bulk of the data are iid $N_4(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = diag(1, 2, 3, 4)$. The first pair of matrices used $\gamma = 0$. Here the FCH, RFCH, and RMVN estimators are $\sqrt{n}$ consistent estimators of $\boldsymbol{\Sigma}$, while $\boldsymbol{C}_{FMCD}$ seems to be approximately unbiased for $0.94\boldsymbol{\Sigma}$.

Next the data had $\gamma = 0.4$ and the outliers had $\boldsymbol{x} \sim N_4((0, 0, 0, 15)^T$, $0.0001\ \boldsymbol{I}_4)$, a near point mass at the major axis. FCH and RFCH estimated $1.93\boldsymbol{\Sigma}$ while RMVN estimated $\boldsymbol{\Sigma}$. The FMCD estimator failed to estimate $d\ \boldsymbol{\Sigma}$. Note that $\chi^2_{4,5/6}/\chi^2_{4,0.5} = 1.9276$.

```
RMVN                                FMCD
 0.988 -0.023 -0.007  0.021   0.227 -0.016  0.002 0.049
-0.023  1.964 -0.022 -0.002  -0.016  0.435 -0.014 0.013
-0.007 -0.022  3.053  0.007   0.002 -0.014  0.673 0.179
 0.021 -0.002  0.007  3.870   0.049  0.013  0.179 55.65
```

Next the data had $\gamma = 0.4$ and the outliers had $\boldsymbol{x} \sim N_4(15\ \boldsymbol{1}, \boldsymbol{\Sigma})$, a mean shift with the same covariance matrix as the clean cases. Again FCH and RFCH estimated $1.93\boldsymbol{\Sigma}$ while RMVN and FMCD estimated $\boldsymbol{\Sigma}$.

```
RMVN                                FMCD
 1.013  0.008  0.006 -0.026   1.024  0.002  0.003 -0.025
 0.008  1.975 -0.022 -0.016   0.002  2.000 -0.034 -0.017
 0.006 -0.022  2.870  0.004   0.003 -0.034  2.931  0.005
-0.026 -0.016  0.004  3.976  -0.025 -0.017  0.005  4.046
```

If $W_{in} \sim N(0, \tau^2/n)$ for $i = 1, ..., r$ and if S^2_W is the sample variance of the W_{in}, then $E(nS^2_W) = \tau^2$ and $V(nS^2_W) = 2\tau^4/(r-1)$. So $nS^2_W \pm \sqrt{5}SE(nS^2_W) \approx \tau^2 \pm \sqrt{10}\tau^2/\sqrt{r-1}$. So for $r = 1000$ runs, we expect nS^2_W to be between $\tau^2 - 0.1\tau^2$ and $\tau^2 + 0.1\tau^2$ with high confidence. Similar results hold for many estimators if W_{in} is $\sqrt{n}$ consistent and asymptotically normal and if n is large enough. If W_{in} has less than $\sqrt{n}$ rate, e.g. $n^{1/3}$ rate, then the scaled sample variance $nS^2_W \to \infty$ as $n \to \infty$.

Table 3.4 considers $W = T_p$ and $W = C_{p,p}$ for eight estimators, $p = 5$ and 10, and $n = 10p$ and 5000, when $\boldsymbol{x} \sim N_p(\boldsymbol{0}, diag(1, ..., p))$. For the classical estimator, denoted by CLAS, $T_p = \overline{x}_p \sim N(0, p/n)$, and $nS^2(T_p) \approx p$ while $C_{p,p}$ is the sample variance of n iid $N(0, p)$ random variables. Hence $nS^2(C_{p,p}) \approx 2p^2$. RFCH, RMVN, FMCD, and OGK use a "reweight for

Table 3.4 Scaled Variance $nS^2(T_p)$ and $nS^2(C_{p,p})$

p	n	V	FCH	RFCH	RMVN	DGK	OGK	CLAS	FMCD	MB
5	50	C	216.0	72.4	75.1	209.3	55.8	47.12	153.9	145.8
5	50	T	12.14	6.50	6.88	10.56	6.70	4.83	8.38	13.23
5	5000	C	307.6	64.1	68.6	325.7	59.3	48.5	60.4	309.5
5	5000	T	18.6	5.34	5.33	19.33	6.61	4.98	5.40	20.20
10	100	C	817.3	276.4	286.0	725.4	229.5	198.9	459.6	610.4
10	100	T	21.40	11.42	11.68	20.13	12.75	9.69	14.05	24.13
10	5000	C	955.5	237.9	243.8	966.2	235.8	202.4	233.6	975.0
10	5000	T	29.12	10.08	10.09	29.35	12.81	9.48	10.06	30.20

efficiency" concentration step that uses a random number of cases with percentage close to 97.5%. These four estimators had similar behavior. DGK, FCH, and MB used about 50% of the cases and had similar behavior. By Lopuhaä (1999), estimators with less than $\sqrt{n}$ rate still have zero efficiency after the reweighting. Although FMCD, MB, and OGK have not been proven to be $\sqrt{n}$ consistent, their values did not blow up even for $n = 5000$.

Geometrical arguments suggest that the MB estimator has considerable outlier resistance. Suppose the outliers are far from the bulk of the data. Let the "median ball" correspond to the half set of data closest to MED($\boldsymbol{W}$) in Euclidean distance. If the outliers are outside of the median ball, then the initial half set in the iteration leading to the MB estimator will be clean. Thus the MB estimator will tend to give the outliers the largest MB distances unless the initial clean half set has very high correlation in a direction about which the outliers lie. This property holds for very general outlier configurations. The FCH estimator tries to use the DGK attractor if the $det(\boldsymbol{C}_{DGK})$ is small and the DGK location estimator T_{DGK} is in the median ball. Distant outliers that make $det(\boldsymbol{C}_{DGK})$ small also drag T_{DGK} outside of the median ball. Then FCH uses the MB attractor.

Compared to OGK and FMCD, the MB estimator is vulnerable to outliers that lie within the median ball. If the bulk of the data is highly correlated with the major axis of a hyperellipsoidal region, then the distances based on the clean data can be very large for outliers that fall within the median ball. The outlier resistance of the MB estimator decreases as p increases since the volume of the median ball rapidly increases with p.

A simple simulation for outlier resistance is to count the number of times the minimum distance of the outliers is larger than the maximum distance of the clean cases. The simulation used 100 runs. If the count was 97, then in 97 data sets the outliers can be separated from the clean cases with a horizontal line in the DD plot, but in 3 data sets the robust distances did not achieve complete separation. In Spring 2015, Det-MCD simulated much like FMCD, but was more likely to cause an error in *R*.

The clean cases had $\boldsymbol{x} \sim N_p(\boldsymbol{0}, diag(1, 2, ..., p))$. Outlier types were the mean shift $\boldsymbol{x} \sim N_p(pm\boldsymbol{1}, diag(1, 2, ..., p))$ where $\boldsymbol{1} = (1, ..., 1)^T$ and $\boldsymbol{x} \sim$

$N_p((0,...,0,pm)^T, 0.0001\boldsymbol{I}_p)$, a near point mass at the major axis. Notice that the clean data can be transformed to a $N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ distribution by multiplying $\boldsymbol{x}_i$ by $\text{diag}(1, 1/\sqrt{2}, ..., 1/\sqrt{p})$, and this transformation changes the location of the near point mass to $(0,...,0,pm/\sqrt{p})^T$.

Table 3.5 Number of Times Mean Shift Outliers had the Largest Distances

p	γ	n	pm	MBA	FCH	RFCH	RMVN	OGK	FMCD	MB
10	.1	100	4	49	49	85	84	38	76	57
10	.1	100	5	91	91	99	99	93	98	91
10	.4	100	7	90	90	90	90	0	48	100
40	.1	100	5	3	3	3	3	76	3	17
40	.1	100	8	36	36	37	37	100	49	86
40	.25	100	20	62	62	62	62	100	0	100
40	.4	100	20	20	20	20	20	0	0	100
40	.4	100	35	44	98	98	98	95	0	100
60	.1	200	10	49	49	49	52	100	30	100
60	.1	200	20	97	97	97	97	100	35	100
60	.25	200	25	60	60	60	60	100	0	100
60	.4	200	30	11	21	21	21	17	0	100
60	.4	200	40	21	100	100	100	100	0	100

For near point mass outliers, a hyperellipsoid with very small volume can cover half of the data if the outliers are at one end of the hyperellipsoid and some of the clean data are at the other end. This half set will produce a classical estimator with very small determinant by Theorem 3.10. In the simulations for large γ, as the near point mass is moved very far away from the bulk of the data, only the classical, MB, and OGK estimators did not have numerical difficulties. Since the MCD estimator has smaller determinant than DGK while MVE has smaller volume than DGK, estimators like FMCD and MBA that use the MVE or MCD criterion without using location information will be vulnerable to these outliers. FMCD is also vulnerable to outliers if γ is slightly larger than γ_o given by (3.23).

Table 3.6 Number of Times Near Point Mass Outliers had the Largest Distances

p	γ	n	pm	MBA	FCH	RFCH	RMVN	OGK	FMCD	MB
10	.1	100	40	73	92	92	92	100	95	100
10	.25	100	25	0	99	99	90	0	0	99
10	.4	100	25	0	100	100	100	0	0	100
40	.1	100	80	0	0	0	0	79	0	80
40	.1	100	150	0	65	65	65	100	0	99
40	.25	100	90	0	88	87	87	0	0	88
40	.4	100	90	0	91	91	91	0	0	91
60	.1	200	100	0	0	0	0	13	0	91
60	.25	200	150	0	100	100	100	0	0	100
60	.4	200	150	0	100	100	100	0	0	100
60	.4	200	20000	0	100	100	100	64	0	100

Tables 3.5 and 3.6 help illustrate the results for the simulation. Large counts and small pm for fixed γ suggest greater ability to detect outliers. Values of p were 5, 10, 15, ..., 60. First consider the mean shift outliers and Table 3.5. For $\gamma = 0.25$ and 0.4, MB usually had the highest counts. For $5 \leq p \leq 20$ and the mean shift, the OGK estimator often had the smallest counts, and FMCD could not handle 40% outliers for $p = 20$. For $25 \leq p \leq 60$, OGK usually had the highest counts for $\gamma = 0.05$ and 0.1. For $p \geq 30$, FMCD could not handle 25% outliers even for enormous values of pm.

In Table 3.6, FCH greatly outperformed MBA although the only difference between the two estimators is that FCH uses a location criterion as well as the MCD criterion. OGK performed well for $\gamma = 0.05$ and $20 \leq p \leq 60$ (not tabled). For large γ, OGK often has large bias for $c\boldsymbol{\Sigma}$. Then the outliers may need to be enormous before OGK can detect them. Also see Table 3.2, where OGK gave the outliers the largest distances for all runs, but $\boldsymbol{C}_{OGK}$ does not give a good estimate of $c\boldsymbol{\Sigma} = c \ \ diag(1, 2)$.

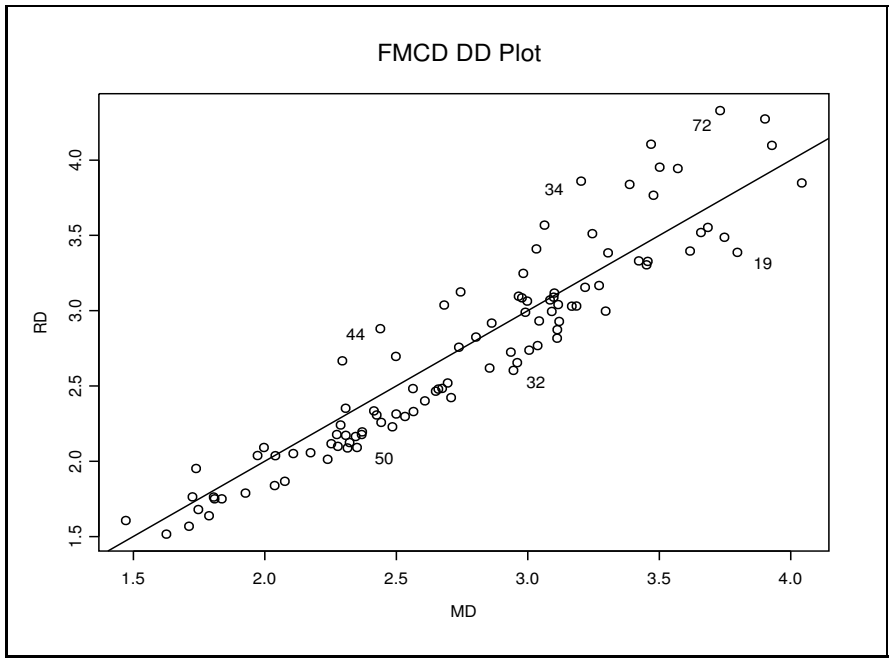

Fig. 3.7 The FMCD Estimator Failed

The DD plot of MD_i versus RD_i is useful for detecting outliers. The resistant estimator will be useful if $(T, \boldsymbol{C}) \approx (\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ where $c > 0$ since scaling by c affects the vertical labels of the RD_i but not the shape of the DD plot. For the outlier data, the MBA estimator is biased, but the mean shift outliers in the MBA DD plot will have large RD_i since $\boldsymbol{C}_{MBA} \approx 2\boldsymbol{C}_{FMCD} \approx 2\boldsymbol{\Sigma}$.

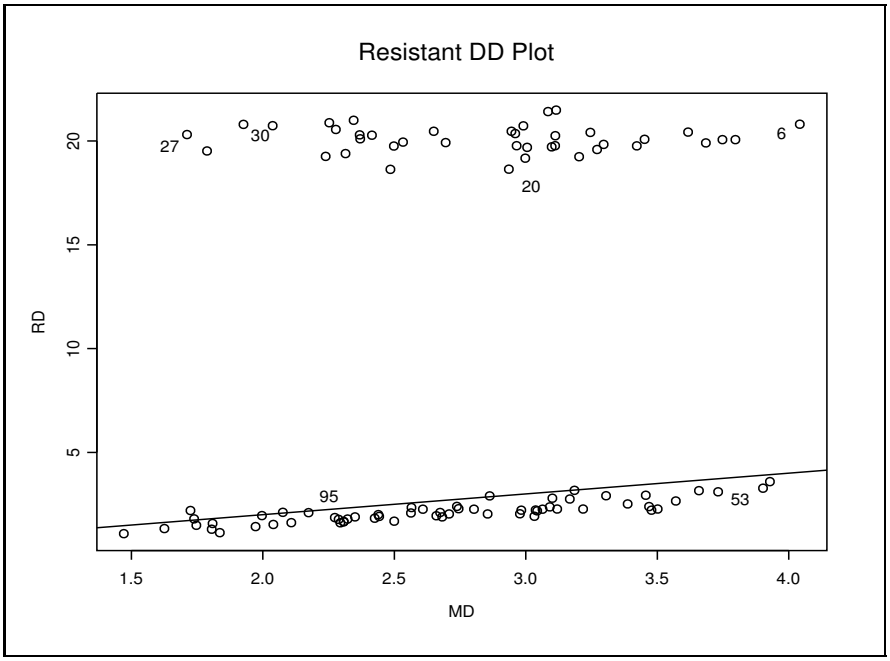

Fig. 3.8 The Outliers are Large in the MBA DD Plot

In an older mean shift simulation, when p was 8 or larger, the `cov.mcd` estimator was usually not useful for detecting the mean shift outliers. Figure 3.7 shows that now the FMCD RD_i are highly correlated with the MD_i. The DD plot based on the MBA estimator detects the outliers. See Figure 3.8.

For many data sets, Equation (3.23) gives a rough approximation for the number of large outliers that concentration algorithms using K starts each consisting of h cases can handle. However, if the data set is multivariate and the bulk of the data falls in one compact hyperellipsoid while the outliers fall in another hugely distant compact hyperellipsoid, then a concentration algorithm using a single start can sometimes tolerate nearly 25% outliers. For example, suppose that all $p + 1$ cases in the elemental start are outliers but the covariance matrix is nonsingular so that the Mahalanobis distances can be computed. Then the classical estimator is applied to the $c_n \approx n/2$ cases with the smallest distances. Suppose the percentage of outliers is less than 25% and that all of the outliers are in this "half set." Then the sample mean applied to the c_n cases should be closer to the bulk of the data than to the cluster of outliers. Hence after a concentration step, the percentage of outliers will be reduced if the outliers are very far away. After the next concentration step the percentage of outliers will be further reduced and after several iterations, all c_n cases will be clean.

In a small simulation study, 20% outliers were planted for various values of p. If the outliers were distant enough, then the minimum DGK distance for the outliers was larger than the maximum DGK distance for the nonoutliers. Hence the outliers would be separated from the bulk of the data in a DD plot of classical versus robust distances. For example, when the clean data comes from the $N_p(\mathbf{0}, \boldsymbol{I}_p)$ distribution and the outliers come from the $N_p(2000\,\mathbf{1}, \boldsymbol{I}_p)$ distribution, the DGK estimator with 10 concentration steps was able to separate the outliers in 17 out of 20 runs when $n = 9000$ and $p = 30$. With 10% outliers, a shift of 40, $n = 600$, and $p = 50$, 18 out of 20 runs worked. Olive (2004a) showed similar results for the Rousseeuw and Van Driessen (1999) FMCD algorithm and that the MBA estimator could often correctly classify up to 49% distant outliers. The following proposition shows that it is very difficult to drive the determinant of the dispersion estimator from a concentration algorithm to zero.

Theorem 3.24. Consider the concentration and MCD estimators that both cover c_n cases. For multivariate data, if at least one of the starts is nonsingular, then the concentration attractor $\boldsymbol{C}_A$ is less likely to be singular than the high breakdown MCD estimator $\boldsymbol{C}_{MCD}$.

Proof. If all of the starts are singular, then the Mahalanobis distances cannot be computed and the classical estimator can not be applied to c_n cases. Suppose that at least one start was nonsingular. Then $\boldsymbol{C}_A$ and $\boldsymbol{C}_{MCD}$ are both sample covariance matrices applied to c_n cases, but by definition $\boldsymbol{C}_{MCD}$ minimizes the determinant of such matrices. Hence $0 \leq \det(\boldsymbol{C}_{MCD}) \leq \det(\boldsymbol{C}_A)$. □

Software

The `robustbase` library was downloaded from (www.r-project.org/#doc). § 11.2 explains how to use the source command to get the `mpack` functions in R and how to download a library from R. Type the commands `library(MASS)` and `library(robustbase)` to compute the FMCD and OGK estimators with the `cov.mcd` and `covOGK` functions. To use Det-MCD instead of FMCD, change

```
out <- covMcd(x)  to out <- covMcd(x,nsamp="deterministic"),
```

but in Spring 2015 this change was more likely to cause errors.

The `rpack` function

mldsim(n=200,p=5,gam=.2,runs=100,outliers=1,pm=15)

can be used to produce Tables 3.1, 3.2, 3.4–3.6. Change outliers to 0 to examine the average of $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$. The function `mldsim6` is similar but does not need the `library` command since it compares the FCH, RFCH, CMVE, RCMVE, MB estimators, and the `covmb2` estimator of Section 3.10. See Olive (2017b) for CMVE and RCMVE. The command

sctplt(n=200,p=10,gam=.2,outliers=3, pm=5)

will make an outlier data set. Then the FCH and MB DD plots are made

(click on the right mouse button and highlight stop to go to the next plot) and then the scatterplot matrix. The scatterplot matrix can be used to determine whether the outliers are hard to detect with bivariate or univariate methods. If $p > 10$ the bivariate plots may be too small.

The function *covsim2* can be modified to show that the `R` implementation of FCH is usually much faster than OGK which is much faster than FMCD. The function *corrsim* can be used to simulate the correlations of robust distances with classical distances. For MVN data, the command
corrsim(n=200,p=20,nruns=100,type=5)
suggests that the correlation of the RFCH distances with the classical distances is about 0.97. Changing *type* to 4 suggests that FCH needs $n = 800$ before the correlation is about 0.97. The function `corrsim2` uses a wider variety of EC distributions.

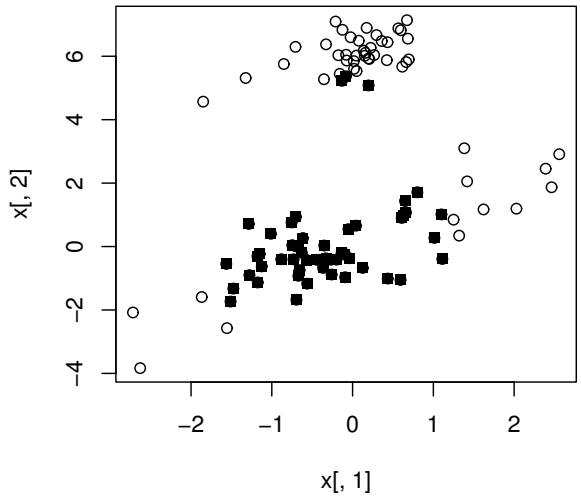

Fig. 3.9 highlighted cases = half set with smallest RD = $(T_0, \boldsymbol{C}_0)$

The function *cmve* computes CMVE and RCMVE, function *covfch* computes FCH and RFCH, while *covrmvn* computes the RMVN and MB estimators. The function *covrmb* computes MB and RMB where RMB is like RMVN except the MB estimator is reweighted instead of FCH. Functions *covdgk*, *covmba*, and *rmba* compute the scaled DGK, MBA, and RMBA estimators. **Better programs would use MB if DGK causes an error.**

The *concmv* function described in Problem 3.30 illustrates concentration where the start is $(\text{MED}(\boldsymbol{W}), diag([MAD(X_i)]^2))$. In Figures 3.9, 3.10, and

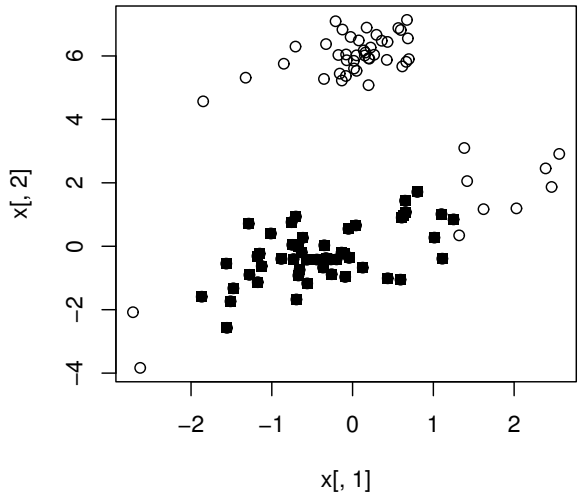

Fig. 3.10 highlighted cases = half set with smallest RD = $(T_1, \boldsymbol{C}_1)$

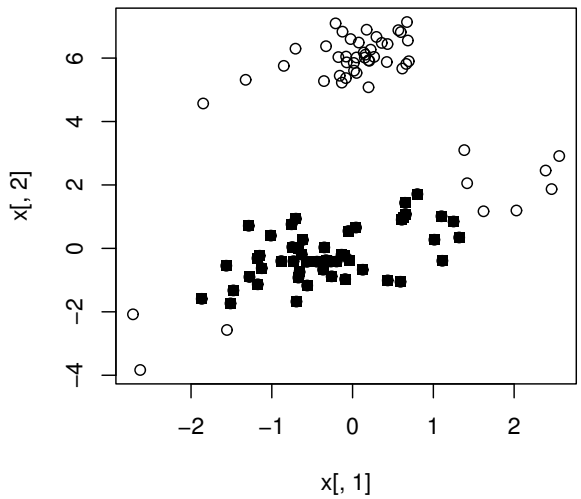

Fig. 3.11 highlighted cases = half set with smallest RD = $(T_2, \boldsymbol{C}_2)$

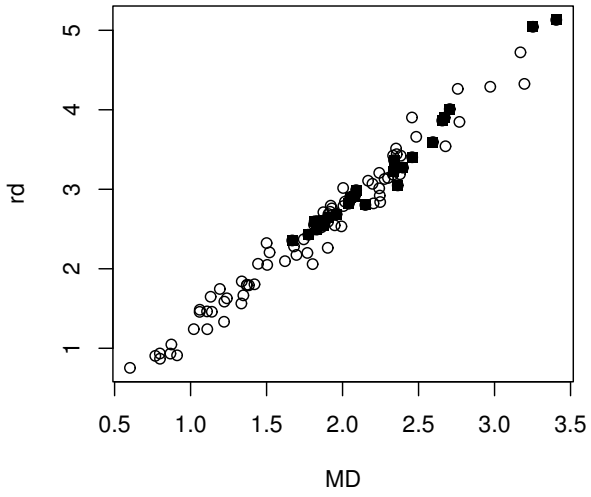

Fig. 3.12 highlighted cases = outliers, RD = $(T_{0,D}, \boldsymbol{C}_{0,D})$

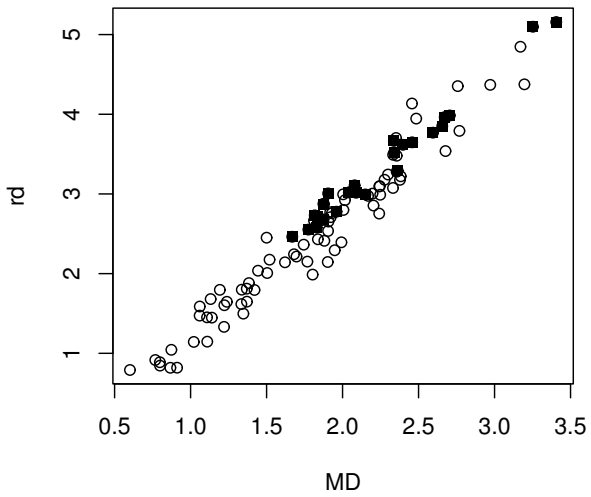

Fig. 3.13 highlighted cases = outliers, RD = $(T_{1,D}, \boldsymbol{C}_{1,D})$

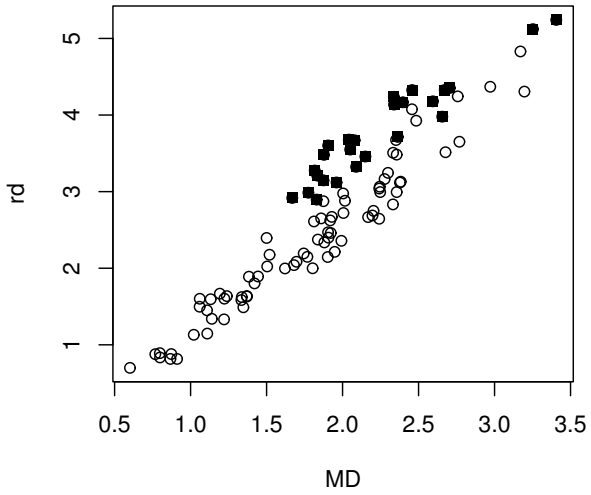

Fig. 3.14 highlighted cases = outliers, RD = $(T_{2,D}, \boldsymbol{C}_{2,D})$

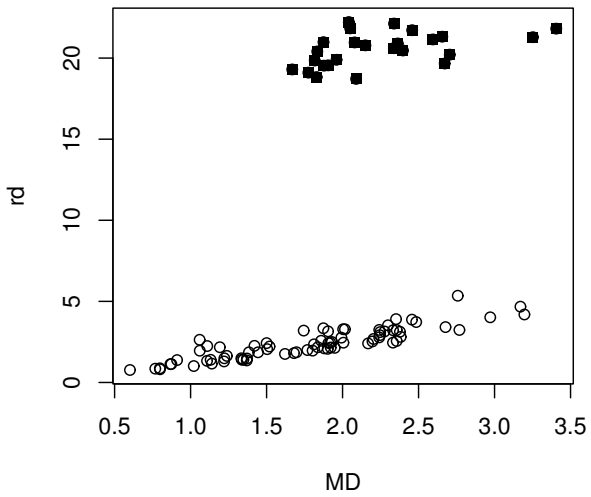

Fig. 3.15 highlighted cases = outliers, RD = $(T_{3,D}, \boldsymbol{C}_{3,D})$

3.11, the highlighted cases are the half set with the smallest distances, and the initial half set shown in Figure 3.9 is not clean, where $n = 100$ and there are 40 outliers. The attractor shown in Figure 3.11 is clean. This type of data set has too many outliers for DGK while the MB starts and attractors are almost always clean.

The *ddmv* function in Problem 3.31 illustrates concentration for the DGK estimator where the start is the classical estimator. Now $n = 100, p = 4$, and there are 25 outliers. A DD plot of classical distances MD versus robust distances RD is shown. See Figures 3.12, 3.13, 3.14, and 3.15. The half set of cases with the smallest RDs is used, and the initial half set shown in Figure 3.12 is not clean. The attractor in Figure 3.15 is the DGK estimator which uses a clean half set. The clean cases $\boldsymbol{x}_i \sim N_4(\mathbf{0}, diag(1, 2, 3, 4))$ while the outliers $\boldsymbol{x}_i \sim N_4((10, 10\sqrt{2}, 10\sqrt{3}, 20)^T,\ diag(1, 2, 3, 4))$.

3.10 Outlier Detection if $p > n$

Most outlier detection methods work best if $n \geq 20p$, but often data sets have $p > n$, and outliers are a major problem. One of the simplest outlier detection methods uses the Euclidean distances of the $\boldsymbol{x}_i$ from the coordinatewise median $D_i = D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$. Concentration type steps compute the weighted median MED_j: the coordinatewise median computed from the "half set" of cases $\boldsymbol{x}_i$ with $D_i^2 \leq \text{MED}(D_i^2(\text{MED}_{j-1}, \boldsymbol{I}_p))$ where $\text{MED}_0 = \text{MED}(\boldsymbol{W})$. We often used $j = 0$ (no concentration type steps) or $j = 9$. Let $D_i = D_i(\text{MED}_j, \boldsymbol{I}_p)$. Let $W_i = 1$ if $D_i \leq \text{MED}(D_1, ..., D_n) + k\text{MAD}(D_1, ..., D_n)$ where $k \geq 0$ and $k = 5$ is the default choice. Let $W_i = 0$, otherwise. Using $k \geq 0$ insures that at least half of the cases get weight 1. This weighting corresponds to the weighting that would be used in a one sided metrically trimmed mean (Huber type skipped mean) of the distances.

Definition 3.26. Let the *covmb2 set* B of at least $n/2$ cases correspond to the cases with weight $W_i = 1$. Then the *covmb2* estimator $(T, \boldsymbol{C})$ is the sample mean and sample covariance matrix applied to the cases in set B. Hence

$$T = \frac{\sum_{i=1}^n W_i \boldsymbol{x}_i}{\sum_{i=1}^n W_i} \text{ and } \boldsymbol{C} = \frac{\sum_{i=1}^n W_i (\boldsymbol{x}_i - T)(\boldsymbol{x}_i - T)^T}{\sum_{i=1}^n W_i \ - \ 1}.$$

Example 3.8. Let the clean data (nonoutliers) be i **1** for $i = 1, 2, 3, 4$, and 5 while the outliers are j **1** for $j = 16, 17, 18$, and 19. Here $n = 9$ and **1** is $p \times 1$. Making a plot of the data for $p = 2$ may be useful. Then the coordinatewise median $\text{MED}_0 = \text{MED}(\boldsymbol{W}) = 5$ **1**. The median Euclidean distance of the data is the Euclidean distance of 5 **1** from 1 **1** = the Euclidean distance of 5 **1** from 9 **1**. The *median ball* is the hypersphere centered at the coordinatewise median with radius $r = \text{MED}(D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p),\ i = 1, ..., n)$ that tends to contain

$(n+1)/2$ of the cases if n is odd. Hence the clean data are in the median ball and the outliers are outside of the median ball. The coordinatewise median of the cases with the 5 smallest distances is the coordinatewise median of the clean data: $\text{MED}_1 = 3\ \mathbf{1}$. Then the median Euclidean distance of the data from MED_1 is the Euclidean distance of $3\ \mathbf{1}$ from $1\ \mathbf{1}$ = the Euclidean distance of $3\ \mathbf{1}$ from $5\ \mathbf{1}$. Again the clean cases are the cases with the 5 smallest Euclidean distances. Hence $\text{MED}_j = 3\ \mathbf{1}$ for $j \geq 1$. For $j \geq 1$, if $\boldsymbol{x}_i = j\ \mathbf{1}$, then $D_i = |j-3|\sqrt{p}$. Thus $D_{(1)} = 0$, $D_{(2)} = D_{(3)} = \sqrt{p}$, and $D_{(4)} = D_{(5)} = 2\sqrt{p}$. Hence $\text{MED}(D_1, ..., D_n) = D_{(5)} = 2\sqrt{p} = \text{MAD}(D_1, ..., D_n)$ since the median distance of the D_i from $D_{(5)}$ is $2\sqrt{p} - 0 = 2\sqrt{p}$. Note that the 5 smallest absolute distances $|D_i - D_{(5)}|$ are $0, 0, \sqrt{p}, \sqrt{p}$, and $2\sqrt{p}$. Hence $W_i = 1$ if $D_i \leq 2\sqrt{p} + 10\sqrt{p} = 12\sqrt{p}$. The clean data get weight 1 while the outliers get weight 0 since the smallest distance D_i for the outliers is the Euclidean distance of $3\ \mathbf{1}$ from $16\ \mathbf{1}$ with a $D_i = \|16\ \mathbf{1} - 3\ \mathbf{1}\| = 13\sqrt{p}$. Hence the `covmb2` estimator $(T, \boldsymbol{C})$ is the sample mean and sample covariance matrix of the clean data. **Note that the distance for the outliers to get zero weight is proportional to the square root of the dimension $\sqrt{p}$.**

The `covmb2` estimator can also be used for $n > p$. The `covmb2` estimator attempts to give a robust dispersion estimator that reduces the bias by using a big ball about MED_j instead of a ball that contains half of the cases. The *rpack* function `getB` gives the set B of cases that got weight 1 along with the index `indx` of the case numbers that got weight 1. The function `ddplot5` plots the Euclidean distances from the coordinatewise median versus the Euclidean distances from the `covmb2` location estimator. Typically the plotted points in this DD plot cluster about the identity line, and outliers appear in the upper right corner of the plot with a gap between the bulk of the data and the outliers. An alternative for outlier detection is to replace $\boldsymbol{C}$ by $\boldsymbol{C}_d = diag(\hat{\sigma}_{11}, ..., \hat{\sigma}_{pp})$. For example, use $\hat{\sigma}_{ii} = \boldsymbol{C}_{ii}$. See Ro et al. (2015) and Tarr et al. (2016) for references.

The next section gives applications of the sets used to compute the RMVN, RFCH, and `covmb2` estimators.

3.11 The RMVN Set, RFCH Set, and covmb2 Set

The RMVN, RFCH, and `covmb2` estimators are each computed from a set of at least $n/2$ cases. We will call these sets the RMVN set U, the RFCH set V and the `covmb2` set B, which was was given in Definition 3.26.

Definition 3.27. Let the n_2 cases in Definition 3.24 be known as the *RMVN set* U. Let the RFCH set V be the set of $m \geq n/2$ cases from which the RFCH estimator is computed.

Referring to Definition 3.24, $(T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2) = (\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ is the classical estimator applied to the RMVN set U, which can be regarded as the untrimmed data (the data not trimmed by ellipsoidal trimming) or the cleaned data. Also $\boldsymbol{S}_U$ is the unscaled estimated dispersion matrix while $\boldsymbol{C}_{RMVN}$ is the scaled estimated dispersion matrix. For the RFCH estimator, $(\overline{\boldsymbol{x}}_V, \boldsymbol{S}_V) = (T_{RFCH}, \tilde{\boldsymbol{\Sigma}}_2)$, and then $\boldsymbol{S}_V$ is scaled to form $\boldsymbol{C}_{RFCH}$.

The two main ways to handle outliers are i) apply the multivariate method to the cleaned data, and ii) plug in robust estimators for classical estimators. Subjectively cleaned data may work well for a single data set, but we can't get large sample theory since sometimes too many cases are deleted (delete outliers and some nonoutliers) and sometimes too few (do not get all of the outliers). Practical plug in robust estimators have rarely been shown to be $\sqrt{n}$ consistent and highly outlier resistant.

Using the RMVN set U or RFCH set V is simultaneously a plug in method and an objective way to clean the data such that the resulting robust method is often backed by theory. Let D be either the set U or V. This result is extremely useful computationally: apply the classical method to the cases in the set D. This procedure is often equivalent to using $(\overline{\boldsymbol{x}}_D, \boldsymbol{S}_D)$ as plug in estimators. The method can be applied if $n > 2(p+1)$ but may not work well unless $n > 20p$. The *rpack* function `getu` gets the RMVN set U as well as the case numbers corresponding to the cases in U. The `covmb2` set B can also be used for several applications, even if $p > n$.

The set D corresponds to a small volume hyperellipsoid containing at least half of the cases since concentration is used. The set D can also be regarded as the "untrimmed data": the data that was not trimmed by ellipsoidal trimming. Theory has been proved for a large class of elliptically contoured distributions, but it is conjectured that theory holds for a much wider class of distributions. See Conjectures 3.3 and 3.4 in Section 3.12. In simulations RFCH and RMVN seem to estimate $c\boldsymbol{\Sigma}_{\boldsymbol{x}}$ if $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{\mu}$ where $\boldsymbol{z} = (z_1, ..., z_p)^T$ and the z_i are iid from a continuous distribution with variance σ^2. Here $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \mathrm{Cov}(\boldsymbol{x}) = \sigma^2 \boldsymbol{A}\boldsymbol{A}^T$. The bias for the MB estimator seemed to be small. It is known that affine equivariant estimators give unbiased estimators of $c\boldsymbol{\Sigma}_{\boldsymbol{x}}$ if the distribution of z_i is also symmetric. DGK is affine equivariant and RFCH and RMVN are asymptotically equivalent to a scaled DGK estimator. But in the simulations the results also held for skewed distributions.

Several illustrative applications are given next, where the theory usually assumes that the cases are iid from a large class of elliptically contoured distributions. There are many other "robust methods" in the literature that use plug in estimators like FMCD. Replacing the plug in estimator by RMVN or RFCH will often greatly improve the robust method.

i) The classical estimator of multivariate location and dispersion applied to the cases in D gives $(\overline{\boldsymbol{x}}_D, \boldsymbol{S}_D)$, a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ for some constant $c > 0$.

ii) The classical estimator of the correlation matrix applied to the cases in U gives $\boldsymbol{R}_U$, a consistent estimator of the population correlation matrix $\boldsymbol{\rho}_{\boldsymbol{x}}$.

iii) For principal component analysis (PCA), RPCA is the classical PCA method applied to the set U. See Olive (2017b, ch. 6).

iv) For canonical correlation analysis (CCA), RCCA is the classical CCA method applied to the set U. See Olive (2017b, ch. 7).

v) Let D_i be the RMVN or RFCH subset applied to the n_i cases from group i for $i = 1, ..., G$. Let $(\overline{\boldsymbol{x}}_{D_i}, \boldsymbol{S}_{D_i})$ be the sample mean and covariance applied to the cases in D_i. Let $Y = i$ for cases in D_i which are from group i. Let $D_{big} = D_1 \cup D_2 \cup \cdots \cup D_G$ be the combined sample. Then apply the discriminant analysis method to D_{big} with the corresponding labels Y. For example, RFDA consists of applying classical FDA on U_{big}. See Olive (2017b, § 8.9).

vi) For factor analysis, apply the factor analysis method to the set D. This method can be used as a diagnostic for methods such as the maximum likelihood method of factor analysis, but is backed by theory for principal component factor analysis. See Olive (2017b, § 11.2).

vii) For multiple linear regression, let Y be the response variable, $x_1 = 1$ and $x_2, ..., x_p$ be the predictor variables. Let $\boldsymbol{z}_i = (Y_i, x_{i2}, ..., x_{ip})^T$. Let D be the RMVN or RFCH set formed using the $\boldsymbol{z}_i$. Then a classical regression estimator applied to the set D results in a robust regression estimator. For least squares, this is implemented with the *rpack* function `rmreg3` using the RMVN set U.

viii) For multivariate linear regression, let $Y_1, ..., Y_m$ be the response variables, $x_1 = 1$ and $x_2, ..., x_p$ be the predictor variables. Let

$$\boldsymbol{z}_i = (Y_{i1}, ...Y_{im}, x_{i2}, ..., x_{ip})^T.$$

Let D be the RMVN or RFCH set formed using the $\boldsymbol{z}_i$. Then a classical least squares multivariate linear regression estimator applied to the set D results in a robust multivariate linear regression estimator. For least squares, this is implemented with the *mpack* function `rmreg3` using U. The method for multiple linear regression in vii) corresponds to $m = 1$. See Olive (2017b, § 12.6.2).

There are also several variants on the method. Suppose there are tentative predictors $Z_1, ..., Z_J$. After transformations assume that predictors $X_1, ..., X_k$ are linearly related. Assume the set U used cases $i_1, i_2, ..., i_{n_U}$. To add variables like $X_{k+1} = X_1^2$, $X_{k+2} = X_3X_4$, $X_{k+3} = gender$, ..., X_p, augment U with the variables $X_{k+1}, ..., X_p$ corresponding to cases $i_1, ..., i_{n_U}$. Adding variables results in cleaned data that is more likely to contain outliers.

If there are g groups ($g = G$ for discriminant analysis, $g = 2$ for binary regression, and $g = p$ for one way MANOVA), the function `getubig` gets the RMVN set U_i for each group and combines the g RMVN sets into one large set $U_{big} = U_1 \cup U_2 \cup \cdots \cup U_g$.

Application 3.3. This outlier resistant regression method uses terms from the following definition. Let the ith case $\boldsymbol{w}_i = (Y_i, \boldsymbol{x}_i^T)^T$ where the continuous predictors from $\boldsymbol{x}_i$ are denoted by $\boldsymbol{u}_i$ for $i = 1, ..., n$. Now let D be the RMVN set U, the RFCH set V or the `covmb2` set B. Find D by applying the estimator to the $\boldsymbol{u}_i$, and then run the regression method on the m cases $\boldsymbol{w}_i$ corresponding to the set D indices $i_1, ..., i_m$, where $m \geq n/2$. The set B can be used even if $p > n$. A similar technique can be used for multivariate regression where the ith case $\boldsymbol{w}_i = (\boldsymbol{y}_i^T, \boldsymbol{x}_i^T)^T$ where the response vector $\boldsymbol{y}_i = (Y_{i1}, ..., Y_{im})^T$ has $m \geq 1$ response variables.

Example 3.9. For the Buxton (1920) data with multiple linear regression, *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! See Problem 3.42 to reproduce the following plots.

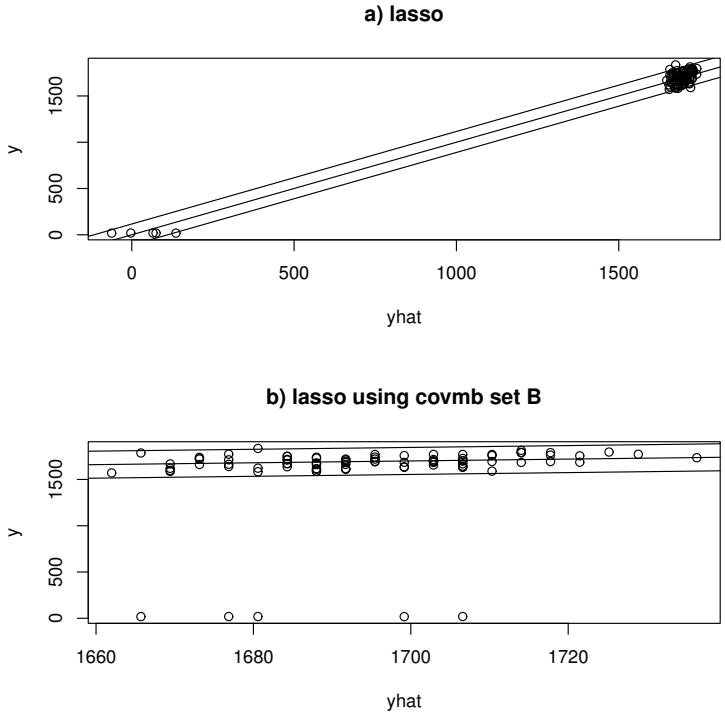

Fig. 3.16 Response plot for lasso and lasso applied to the `covmb2` set B.

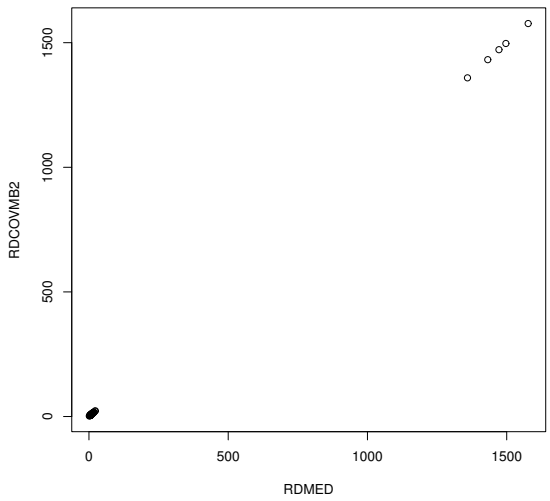

Fig. 3.17 DD plot.

Figure 3.16a) shows the response plot for lasso. The identity line passes right through the outliers which are obvious because of the large gap. Figure 3.16b) shows the response plot from lasso for the cases in the `covmb2` set B applied to the predictors, and the set B included all of the clean cases and omitted the 5 outliers. The response plot was made for all of the data, including the outliers. Prediction interval (PI) bands are also included for both plots. Both plots are useful for outlier detection, but the method for plot 3.16b) is better for data analysis: impossible outliers should be deleted or given 0 weight, we do not want to predict that some people are about 0.75 inches tall, and we do want to predict that the people were about 1.6 to 1.8 meters tall. Figure 3.17 shows the DD plot made using `ddplot5`. The five outliers are in the upper right corner.

The *rpack* function `mldsim6` suggests that for 40% outliers, the outliers need to be further away from the bulk of the data for `covmb2` (`covmb2(k=5)` needs a larger value of pm) than for the other six estimators if $n \geq 20p$. With some outlier types, `covmb2(k=5)` was often near best. Try the following commands. The other estimators need $n > 2p$, and as n gets close to $2p$, `covmb2` may outperform the other estimators.

```
#near point mass on major axis
mldsim6(n=100,p=10,outliers=1,gam=0.25,pm=25)
mldsim6(n=100,p=10,outliers=1,gam=0.4,pm=25) #bad
mldsim6(n=100,p=40,outliers=1,gam=0.1,pm=100)
```

```
mldsim6(n=200,p=60,outliers=1,gam=0.1,pm=100)
#mean shift outliers
mldsim6(n=100,p=40,outliers=3,gam=0.1,pm=10)
mldsim6(n=100,p=40,outliers=3,gam=0.25,pm=20)
mldsim6(n=200,p=60,outliers=3,gam=0.1,pm=10)
#concentration steps can help
mldsim6(n=100,p=10,outliers=3,gam=0.4,pm=10,osteps=0)
mldsim6(n=100,p=10,outliers=3,gam=0.4,pm=10,osteps=9)
```

3.12 Summary

The following three quantities are important.

1) $E(\boldsymbol{x}) = \boldsymbol{\mu} = (E(x_1), ..., E(x_p))^T$.

2) The $p \times p$ *population covariance matrix* $\mathrm{Cov}(\boldsymbol{x}) = E(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T = (\sigma_{ij}) = \boldsymbol{\Sigma_x}$.

3) The $p \times p$ *population correlation matrix* $\mathrm{Cor}(\boldsymbol{x}) = \boldsymbol{\rho_x} = (\rho_{ij})$.

4) The *population covariance matrix* of $\boldsymbol{x}$ with $\boldsymbol{y}$ is $\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\Sigma_{x,y}} = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{y} - E(\boldsymbol{y}))^T]$.

5) Let the $p \times p$ matrix $\boldsymbol{\Delta} = \mathrm{diag}(\sqrt{\sigma_{11}}, ..., \sqrt{\sigma_{\mathrm{pp}}})$. Then $\boldsymbol{\Sigma_x} = \boldsymbol{\Delta \rho_x \Delta}$, and $\boldsymbol{\rho_x} = \boldsymbol{\Delta}^{-1} \boldsymbol{\Sigma_x} \boldsymbol{\Delta}^{-1}$.

6) The $n \times p$ data matrix

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_p \end{bmatrix}.$$

7) The **sample mean** or *sample mean vector*

$$\overline{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i = (\overline{x}_1, ..., \overline{x}_p)^T = \frac{1}{n} \boldsymbol{W}^T \boldsymbol{1}$$

where $\boldsymbol{1}$ is the $p \times 1$ vector of ones.

8) The **sample covariance matrix**

$$\boldsymbol{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T = (S_{ij}).$$

9) The classical estimator of multivariate location and dispersion is $(\overline{\boldsymbol{x}}, \boldsymbol{S})$.

10) $(n-1)\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T - \overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^T = (\boldsymbol{W} - \boldsymbol{1}\overline{\boldsymbol{x}}^T)^T(\boldsymbol{W} - \boldsymbol{1}\overline{\boldsymbol{x}}^T) =$
$\boldsymbol{W}^T\boldsymbol{W} - \frac{1}{n}\boldsymbol{W}^T\boldsymbol{1}\boldsymbol{1}^T\boldsymbol{W}$. Hence if the *centering matrix* $\boldsymbol{H} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T$, then $(n-1)\boldsymbol{S} = \boldsymbol{W}^T\boldsymbol{H}\boldsymbol{W}$.

11) The **sample correlation matrix** $\boldsymbol{R} = (r_{ij})$.

12) Let the $p \times p$ sample standard deviation matrix $\boldsymbol{D} = \text{diag}(\sqrt{\text{S}_{11}}, ..., \sqrt{\text{S}_{\text{pp}}})$. Then $\boldsymbol{S} = \boldsymbol{D}\boldsymbol{R}\boldsymbol{D}$, and $\boldsymbol{R} = \boldsymbol{D}^{-1}\boldsymbol{S}\boldsymbol{D}^{-1}$.

13) The spectral decomposition of the symmetric matrix $\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T$ $= \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1^T + \cdots + \lambda_p \boldsymbol{e}_p \boldsymbol{e}_p^T$.

14) Let $\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T$ be a positive definite $p \times p$ symmetric matrix. Let $\boldsymbol{P} = [\boldsymbol{e}_1\ \boldsymbol{e}_2\ \cdots\ \boldsymbol{e}_p]$ be the $p \times p$ orthogonal matrix with ith column $\boldsymbol{e}_i$. Let $\boldsymbol{\Lambda}^{1/2} = \text{diag}(\sqrt{\lambda_1}, ..., \sqrt{\lambda_p})$. The *square root matrix* $\boldsymbol{A}^{1/2} = \boldsymbol{P}\boldsymbol{\Lambda}^{1/2}\boldsymbol{P}^T$ is a positive definite symmetric matrix such that $\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2} = \boldsymbol{A}$.

15) The *generalized sample variance* $= |\boldsymbol{S}| = \det(\boldsymbol{S})$.

16) The hyperellipsoid $\{\boldsymbol{x} | D_{\boldsymbol{x}}^2 \le h^2\} = \{\boldsymbol{x} : (\boldsymbol{x} - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x} - \overline{\boldsymbol{x}}) \le h^2\}$ is centered at $\overline{\boldsymbol{x}}$ and has volume equal to

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} |\boldsymbol{S}|^{1/2} h^p.$$

Let $\boldsymbol{S}$ have eigenvalue eigenvector pairs $(\hat{\lambda}_i, \hat{\boldsymbol{e}}_i)$ where $\hat{\lambda}_1 \ge \cdots \ge \hat{\lambda}_p$. If $\overline{\boldsymbol{x}} = \boldsymbol{0}$, the axes are given by the eigenvectors $\hat{\boldsymbol{e}}_i$ where the half length in the direction of $\hat{\boldsymbol{e}}_i$ is $h\sqrt{\hat{\lambda}_i}$. Here $\hat{\boldsymbol{e}}_i^T \hat{\boldsymbol{e}}_j = 0$ for $i \ne j$ while $\hat{\boldsymbol{e}}_i^T \hat{\boldsymbol{e}}_i = 1$.

17) Given a table of data $\boldsymbol{W}$ for variables $X_1, ..., X_p$, be able to find the **coordinatewise median** MED($\boldsymbol{W}$) and the **sample mean** $\overline{\boldsymbol{x}}$. If $\boldsymbol{x} = (X_1, X_2, ..., X_p)^T$ where X_j corresponds to the jth column of $\boldsymbol{W}$, then $\text{MED}(\boldsymbol{W}) = (\text{MED}_{X_1}(n), ..., \text{MED}_{X_p}(n))^T$ where $\text{MED}_{X_j}(n) = \text{MED}(X_{j,1}, ..., X_{j,n})$ is the sample median of the data in the jth column. Similarly, $\overline{\boldsymbol{x}} = (\overline{X}_1, ..., \overline{X}_p)^T$ where $\overline{X}_j$ is the sample mean of the data in the jth column.

18) If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $p \times 1$ random vectors, $\boldsymbol{a}$ a conformable constant vector, and $\boldsymbol{A}$ and $\boldsymbol{B}$ are conformable constant matrices, then

$$E(\boldsymbol{X}+\boldsymbol{Y}) = E(\boldsymbol{X})+E(\boldsymbol{Y}),\ E(\boldsymbol{a}+\boldsymbol{Y}) = \boldsymbol{a}+E(\boldsymbol{Y}),\ \&\ E(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}) = \boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}.$$

Also

$$\text{Cov}(\boldsymbol{a} + \boldsymbol{A}\boldsymbol{X}) = \text{Cov}(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}\text{Cov}(\boldsymbol{X})\boldsymbol{A}^T.$$

Note that $E(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{A}E(\boldsymbol{Y})$ and $\text{Cov}(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{A}\text{Cov}(\boldsymbol{Y})\boldsymbol{A}^T$.

19) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$.

20) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and if $\boldsymbol{A}$ is a $q \times p$ matrix, then $\boldsymbol{A}\boldsymbol{X} \sim N_q(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. If $\boldsymbol{a}$ is a $p \times 1$ vector of constants, then $\boldsymbol{X} + \boldsymbol{a} \sim N_p(\boldsymbol{\mu} + \boldsymbol{a}, \boldsymbol{\Sigma})$.

$$\text{Let } \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix},\ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix},\ \text{and } \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11}\ \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21}\ \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

21) **All subsets of a MVN are MVN:** $(X_{k_1}, ..., X_{k_q})^T \sim N_q(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ where $\tilde{\boldsymbol{\mu}}_i = E(X_{k_i})$ and $\tilde{\boldsymbol{\Sigma}}_{ij} = \mathrm{Cov}(X_{k_i}, X_{k_j})$. In particular, $\boldsymbol{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\boldsymbol{X}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$. If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent iff $\boldsymbol{\Sigma}_{12} = \boldsymbol{0}$.

22)

$$\text{Let} \quad \begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \mathrm{Cov}(Y,X) \\ \mathrm{Cov}(X,Y) & \sigma_X^2 \end{pmatrix} \right).$$

Also recall that the *population correlation* between X and Y is given by

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{VAR}(X)}\sqrt{\mathrm{VAR}(Y)}} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

if $\sigma_X > 0$ and $\sigma_Y > 0$.

23) The conditional distribution of a MVN is MVN. If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal with mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. That is,

$$\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2 \sim N_q(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

24) Notation:

$$\boldsymbol{X}_1 | \boldsymbol{X}_2 \sim N_q(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{X}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

25) Be able to compute the above quantities if X_1 and X_2 are scalars.

26) A $p \times 1$ random vector $\boldsymbol{X}$ has an *elliptically contoured distribution,* if $\boldsymbol{X}$ has joint pdf

$$f(\boldsymbol{z}) = k_p |\boldsymbol{\Sigma}|^{-1/2} g[(\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu})], \tag{3.35}$$

and we say $\boldsymbol{X}$ has an elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution. If the second moments exist, then

$$E(\boldsymbol{X}) = \boldsymbol{\mu} \tag{3.36}$$

and

$$\mathrm{Cov}(\boldsymbol{X}) = c_X \boldsymbol{\Sigma} \tag{3.37}$$

for some constant $c_X > 0$.

27) The *population squared Mahalanobis distance*

$$U \equiv D^2 = D_{\boldsymbol{x}}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}). \tag{3.38}$$

For elliptically contoured distributions, U has pdf

$$h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)} k_p u^{p/2-1} g(u). \tag{3.39}$$

$U \sim \chi^2_p$ if $\boldsymbol{x}$ has a multivariate normal $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution.

29) Let the $p \times 1$ column vector $T(\boldsymbol{W})$ be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C}(\boldsymbol{W})$ be a dispersion estimator. Then the ith *squared sample Mahalanobis distance* is the scalar

$$D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{x}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T(\boldsymbol{W})) \quad (3.40)$$

for each observation $\boldsymbol{x}_i$. Notice that the Euclidean distance of $\boldsymbol{x}_i$ from the estimate of center $T(\boldsymbol{W})$ is $D_i(T(\boldsymbol{W}), \boldsymbol{I}_p)$. The classical Mahalanobis distance uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. Note that $D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}})$.

30) A **DD plot** is a plot of classical vs. robust Mahalanobis distances. The DD plot is used to check i) if the data is MVN (plotted points follow the identity line), ii) if the data is EC but not MVN (plotted points follow a line through the origin with slope > 1), iii) if the data is not EC (plotted points do not follow a line through the origin), iv) if multivariate outliers are present (e.g. some plotted points are far from the bulk of the data or the plotted points follow two lines). v) The DD plot can be used to display the prediction regions of Chapter 4.

31) Many practical "robust estimators" generate a sequence of K trial fits called *attractors*: $(T_1, \boldsymbol{C}_1), ..., (T_K, \boldsymbol{C}_K)$. Then the attractor $(T_A, \boldsymbol{C}_A)$ that minimizes some criterion is used to obtain the final estimator. One way to obtain attractors is to generate trial fits called *starts*, and then use the *concentration* technique. Let $(T_{-1,j}, \boldsymbol{C}_{-1,j})$ be the jth start and compute all n Mahalanobis distances $D_i(T_{-1,j}, \boldsymbol{C}_{-1,j})$. At the next iteration, the classical estimator $(T_{0,j}, \boldsymbol{C}_{0,j})$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest distances. This iteration can be continued for k steps resulting in the sequence of estimators $(T_{-1,j}, \boldsymbol{C}_{-1,j}), (T_{0,j}, \boldsymbol{C}_{0,j}), ..., (T_{k,j}, \boldsymbol{C}_{k,j})$. Then $(T_{k,j}, \boldsymbol{C}_{k,j})$ is the jth attractor for $j = 1, ..., K$. Using $k = 10$ often works well, and the basic resampling algorithm is a special case $k = -1$ where the attractors are the starts.

32) The DGK estimator $(T_{DGK}, \boldsymbol{C}_{DGK})$ uses the classical estimator $(T_{-1,D}, \boldsymbol{C}_{-1,D}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ as the only start.

33) The median ball (MB) estimator $(T_{MB}, \boldsymbol{C}_{MB})$ uses $(T_{-1,M}, \boldsymbol{C}_{-1,M}) = (\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ as the only start where $\text{MED}(\boldsymbol{W})$ is the coordinatewise median. Hence $(T_{0,M}, \boldsymbol{C}_{0,M})$ is the classical estimator applied to the "half set" of data closest to $\text{MED}(\boldsymbol{W})$ in Euclidean distance.

34) Elemental concentration algorithms use elemental starts: $(T_{-1,j}, \boldsymbol{C}_{-1,j}) = (\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ is the classical estimator applied to a randomly selected "elemental set" of $p + 1$ cases. If the $\boldsymbol{x}_i$ are iid with covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, then the starts $(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ are identically distributed with $E(\overline{\boldsymbol{x}}_j) = E(\boldsymbol{x}_i)$, $\text{Cov}(\overline{\boldsymbol{x}}_j) = \boldsymbol{\Sigma}_{\boldsymbol{x}}/(p+1)$, and $E(\boldsymbol{S}_j) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$.

35) Let the "median ball" be the hypersphere containing the half set of data closest to $\text{MED}(\boldsymbol{W})$ in Euclidean distance. The FCH estimator uses the MB attractor if the DGK location estimator $T_{DGK} = T_{k,D}$ is outside of the median ball, and the attractor with the smallest determinant, otherwise. Let

$(T_A, \boldsymbol{C}_A)$ be the attractor used. Then the estimator $(T_{FCH}, \boldsymbol{C}_{FCH})$ takes $T_{FCH} = T_A$ and

$$\boldsymbol{C}_{FCH} = \frac{\text{MED}(D_i^2(T_A, \boldsymbol{C}_A))}{\chi^2_{p,0.5}} \boldsymbol{C}_A \tag{3.41}$$

where $\chi^2_{p,0.5}$ is the 50th percentile of a chi–square distribution with p degrees of freedom. The RFCH estimator uses two standard "reweight for efficiency steps" while the RMVN estimator uses a modified method for reweighting.

36) For a large class of elliptically contoured distributions, FCH, RFCH, and RMVN are $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, c_i \boldsymbol{\Sigma})$ for $c_1, c_2, c_3 > 0$ where $c_i = 1$ for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ data.

37) An estimator $(T, \boldsymbol{C})$ of multivariate location and dispersion (MLD), needs to estimate $p(p+3)/2$ unknown parameters when there are p random variables. For $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ or $(\overline{\boldsymbol{z}}, \boldsymbol{R})$, we want $n \geq 10p$. We want $n \geq 20p$ for FCH, RFCH, or RMVN.

38) Brand name robust MLD estimators take too long to compute: F-brand name estimators that are not backed by breakdown or large sample theory are actually used. FMCD, F-MVE, F-S, F-MM, F-τ, F-constrained-M and F-Stahel-Donoho are especially common. F-brand name estimators use a fixed number of starts.

39) The squared Euclidean distances of the $\boldsymbol{x}_i$ from the coordinatewise median are $D_i^2 = D_i^2(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$. Concentration type steps compute the weighted median MED_j: the coordinatewise median computed from the cases $\boldsymbol{x}_i$ with $D_i^2 \leq \text{MED}(D_i^2(\text{MED}_{j-1}, \boldsymbol{I}_p))$ where $\text{MED}_0 = \text{MED}(\boldsymbol{W})$. Often used $j = 0$ (no concentration type steps) or $j = 9$. Let $D_i = D_i(\text{MED}_j, \boldsymbol{I}_p)$. Let $W_i = 1$ if $D_i \leq \text{MED}(D_1, ..., D_n) + k\text{MAD}(D_1, ..., D_n)$ where $k \geq 0$ and $k = 5$ is the default choice. Let $W_i = 0$, otherwise.

40) Let the *covmb2 set* B of at least $n/2$ cases correspond to the cases with weight $W_i = 1$. Then the *covmb2* estimator $(T, \boldsymbol{C})$ is the sample mean and sample covariance matrix applied to the cases in set B. Hence

$$T = \frac{\sum_{i=1}^n W_i \boldsymbol{x}_i}{\sum_{i=1}^n W_i} \text{ and } \boldsymbol{C} = \frac{\sum_{i=1}^n W_i (\boldsymbol{x}_i - T)(\boldsymbol{x}_i - T)^T}{\sum_{i=1}^n W_i - 1}.$$

The function `ddplot5` plots the Euclidean distances from the coordinatewise median versus the Euclidean distances from the `covmb2` location estimator. Typically the plotted points in this DD plot cluster about the identity line, and outliers appear in the upper right corner of the plot with a gap between the bulk of the data and the outliers.

3.13 Complements

For concentration algorithms, note that $(T_{t,j}, \boldsymbol{C}_{t,j}) = (\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$ is the classical estimator applied to the "half set" of cases satisfying $\{\boldsymbol{x}_i : D_i^2(\overline{\boldsymbol{x}}_{t-1,j}, \boldsymbol{S}_{t-1,j})$

$\leq D^2_{(c_n)}(\overline{\boldsymbol{x}}_{t-1,j}, \boldsymbol{S}_{t-1,j})\}$ for $t \geq 0$. Hence $(T_{t,j}, \boldsymbol{C}_{t,j})$ is estimating $(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, the population mean and covariance matrix of the truncated distribution covering half of the mass corresponding to $\{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu}_{t-1})^T \boldsymbol{\Sigma}_{t-1}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{t-1}) \leq D^2_{0.5}(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})\}$ where $D^2_{0.5}(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})$ is the population median of the population squared distances $D^2(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})$. Here $(\boldsymbol{\mu}_{-1}, \boldsymbol{\Sigma}_{-1})$ is the population analog of $(T_{-1,j}, \boldsymbol{C}_{-1,j})$.

The DGK estimator $(T_{k,D}, \boldsymbol{C}_{k,D})$ uses the classical estimator $(T_{-1,D}, \boldsymbol{C}_{-1,D})$ $= (\overline{\boldsymbol{x}}, \boldsymbol{S})$ as the only start. Thus $(\boldsymbol{\mu}_{-1,D}, \boldsymbol{\Sigma}_{-1,D})$ is the population mean and covariance matrix. For a large class of elliptically contoured distributions with a nonsingular covariance matrix and for $t \geq 0$, $(\boldsymbol{\mu}_{t,D}, \boldsymbol{\Sigma}_{t,D})$ is the population mean and covariance matrix of the truncated distribution corresponding to the highest density region covering half the mass. Hence $\boldsymbol{\mu}_{t,D} = \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_{t,D} = c\boldsymbol{\Sigma}$ for some $c > 0$. Riani, Atkinson and Cerioli (2009) find the population mean and covariance matrices for such truncated multivariate normal distributions, using results from Tallis (1963).

Conjecture 3.3. The DGK estimator is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}_{k,D}, \boldsymbol{\Sigma}_{k,D})$ under mild conditions.

The median ball (MB) estimator $(T_{k,M}, \boldsymbol{C}_{k,M})$ uses $(T_{-1,M}, \boldsymbol{C}_{-1,M}) =$ $(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ as the only start where $\text{MED}(\boldsymbol{X})$ is the coordinatewise median. Hence $(T_{0,M}, \boldsymbol{C}_{0,M})$ is the classical estimator applied to the "half set" of data closest to $\text{MED}(\boldsymbol{W})$ in Euclidean distance while $(\boldsymbol{\mu}_{0,M}, \boldsymbol{\Sigma}_{0,M})$ is the population mean and covariance matrix of the truncated distribution corresponding to the hypersphere centered at the population median that contains half the mass. For a distribution that is spherical about $\boldsymbol{\mu}$ and for $t \geq 0$, $(\boldsymbol{\mu}_{t,M}, \boldsymbol{\Sigma}_{t,M}) = (\boldsymbol{\mu}, c\boldsymbol{I}_p)$ for some $c > 0$. For nonspherical elliptically contoured distributions, $\boldsymbol{\Sigma}_{t,M} \neq c\boldsymbol{\Sigma}$. However, the bias seems to be small even for $t = 0$, and to get smaller as k increases. If the median ball estimator is iterated to convergence, we do not know whether $\boldsymbol{\Sigma}_{\infty,M} = c\boldsymbol{\Sigma}$.

Conjecture 3.4. The MB estimator is a high breakdown $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}_{k,M}, \boldsymbol{\Sigma}_{k,M})$ under mild conditions. For elliptically contoured distributions, $\boldsymbol{\mu}_{k,M} = \boldsymbol{\mu}$.

Arcones (1995) and Kim (2000) showed that $\overline{\boldsymbol{x}}_{0,M}$ is a HB $\sqrt{n}$ consistent estimator of $\boldsymbol{\mu}$. Olive (2004a) showed that $(\overline{\boldsymbol{x}}_{0,M}, \boldsymbol{S}_{0,M}) = (T_{0,m}, \boldsymbol{C}_{0,m})$ is a high breakdown estimator. If the data distribution is EC but not spherical about $\boldsymbol{\mu}$, then for $k \geq 0$, $\boldsymbol{S}_{k,M} = \boldsymbol{C}_{MB}$ under estimates the major axis and over estimates the minor axis of the highest density region. Concentration reduces but fails to eliminate this bias. Hence the estimated highest density region based on the attractor is "shorter" in the direction of the major axis and "fatter" in the direction of the minor axis than estimated regions based on consistent estimators.

This chapter followed Olive (2017b, §s 2.1,2.2, 2.3, 3.1, 3.2, 5.1, ch. 4) closely. The theory for concentration algorithms is due to Hawkins and Olive (2002) and Olive and Hawkins (2010). The MBA estimator is due to Olive (2004a). The computational and theoretical simplicity of the FCH estimator makes it one of the most useful robust estimators ever proposed. The RFCH

and RMVN estimators takes slightly longer to compute than the FCH estimator, and may have slightly less resistance to outliers. These three estimators appear in Zhang, Olive, and Ye (2012). A good paper for the DD plot is Olive (2002). Olive (2017b) showed that the DD plot of the residuals is useful for MANOVA models and for multivariate linear regression models where the response vector $\boldsymbol{y} = (Y_1, ..., Y_m)^T$.

Rousseeuw (1984) introduced the MCD and the minimum volume ellipsoid MVE(c_n) estimator. For the MVE estimator, $T(\boldsymbol{W})$ is the center of the minimum volume ellipsoid covering c_n of the observations and $\boldsymbol{C}(\boldsymbol{W})$ is determined from the same ellipsoid. T_{MVE} has a cube root rate and the limiting distribution is not Gaussian. See Davies (1992).

Estimators with complexity higher than $O[(n^3 + n^2p + np^2 + p^3)\log(n)]$ take too long to compute and will rarely be used. No practical useful "high breakdown" estimator (with complexity less than $O(n^4)$ for general p) of multivariate location and dispersion has been shown to be both consistent and high breakdown. The FCH, RFCH, and RMVN estimators have the most theory. The OGK, Det-MCD, sign covariance matrix and k-step spatial sign covariance matrix are the leading competitors. See Olive (2017b, pp. 124-125) for more on the sign covariance matrix.

It is possible to compute the MCD and MVE estimators for $p = 4$ and $n = 100$ in a few hours using branch and bound algorithms (like estimators with $O(100^4)$ complexity). See Agulló (1996, 1998) and Pesch (1999). These algorithms take too long if both $p \geq 5$ and $n \geq 100$. Simulations may need $p \leq 2$. Two stage estimators such as the MM estimator, that need an initial high breakdown consistent estimator, take longer to compute than the initial estimator. See Maronna et al. (2006, ch. 6) for descriptions and references.

Several outlier detection methods for $p > n$ have been proposed. It would be interesting to see if any of these methods are competitive with the `covmb2` estimator and Euclidean distances from the coordinatewise median. See Boudt et al. (2020), Ro et al. (2015), Tarr et al. (2016) for references. Filsomer et al. (2008) note that RD_i can be computed without matrix inversion, and that in high dimensions, outliers with different shape than inliers tend to lie in different hyperspheres.

3.14 Problems

3.1*. Suppose that

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim N_4 \left(\begin{pmatrix} 49 \\ 100 \\ 17 \\ 7 \end{pmatrix}, \begin{pmatrix} 3 & 1 & -1 & 0 \\ 1 & 6 & 1 & -1 \\ -1 & 1 & 4 & 0 \\ 0 & -1 & 0 & 2 \end{pmatrix} \right).$$

a) Find the distribution of X_2.

b) Find the distribution of $(X_1, X_3)^T$.

c) Which pairs of random variables X_i and X_j are independent?

d) Find the correlation $\rho(X_1, X_3)$.

3.2*. Recall that if $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal with mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$.

Let $\sigma_{12} = \text{Cov}(Y, X)$ and suppose Y and X follow a bivariate normal distribution

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2\left(\begin{pmatrix} 49 \\ 100 \end{pmatrix}, \begin{pmatrix} 16 & \sigma_{12} \\ \sigma_{12} & 25 \end{pmatrix} \right).$$

a) If $\sigma_{12} = 0$, find $Y|X$. Explain your reasoning.

b) If $\sigma_{12} = 10$ find $E(Y|X)$.

c) If $\sigma_{12} = 10$, find $\text{Var}(Y|X)$.

3.3. Let $\sigma_{12} = \text{Cov}(Y, X)$ and suppose Y and X follow a bivariate normal distribution

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2\left(\begin{pmatrix} 15 \\ 20 \end{pmatrix}, \begin{pmatrix} 64 & \sigma_{12} \\ \sigma_{12} & 81 \end{pmatrix} \right).$$

a) If $\sigma_{12} = 10$ find $E(Y|X)$.

b) If $\sigma_{12} = 10$, find $\text{Var}(Y|X)$.

c) If $\sigma_{12} = 10$, find $\rho(Y, X)$, the correlation between Y and X.

3.4. Suppose that

$$\boldsymbol{X} \sim (1 - \gamma)EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g_1) + \gamma EC_p(\boldsymbol{\mu}, c\boldsymbol{\Sigma}, g_2)$$

where $c > 0$ and $0 < \gamma < 1$. Following Example 3.2, show that $\boldsymbol{X}$ has an elliptically contoured distribution assuming that all relevant expectations exist.

3.5. In Theorem 3.5b, show that if the second moments exist, then $\boldsymbol{\Sigma}$ can be replaced by $\text{Cov}(\boldsymbol{X})$.

```
crancap     hdlen   hdht   Data for 3.6
 1485       175     132
 1450       191     117
 1460       186     122
 1425       191     125
 1430       178     120
 1290       180     117
   90        75      51
```

3.6*. The table ($\boldsymbol{W}$) above represents 3 head measurements on 6 people and one ape. Let X_1 = *cranial capacity*, X_2 = *head length* and X_3 = *head height*. Let $\boldsymbol{x} = (X_1, X_2, X_3)^T$. Several multivariate location estimators, including the coordinatewise median and sample mean, are found by applying a univariate location estimator to each random variable and then collecting the results into a vector. a) Find the coordinatewise median MED($\boldsymbol{W}$).

b) Find the sample mean $\overline{\boldsymbol{x}}$.

3.7. Using the notation in Theorem 3.6, show that if the second moments exist, then

$$\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY} = [\mathrm{Cov}(\boldsymbol{X})]^{-1}\mathrm{Cov}(\boldsymbol{X}, Y).$$

3.8. Using the notation under Theorem 3.4, show that if $\boldsymbol{X}$ is elliptically contoured, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is also elliptically contoured.

3.9*. Suppose $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$. Find the distribution of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ if $\boldsymbol{X}$ is an $n \times p$ full rank constant matrix.

3.10. Recall that $\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = E[(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{Y} - E(\boldsymbol{Y}))^T]$. Using the notation of Theorem 3.6, let $(Y, \boldsymbol{X}^T)^T$ be $EC_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ where Y is a random variable. Let the covariance matrix of $(Y, \boldsymbol{X}^T)$ be

$$\mathrm{Cov}((Y, \boldsymbol{X}^T)^T) = c\begin{pmatrix} \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XX} \end{pmatrix} = \begin{pmatrix} \mathrm{VAR}(Y) & \mathrm{Cov}(Y, \boldsymbol{X}) \\ \mathrm{Cov}(\boldsymbol{X}, Y) & \mathrm{Cov}(\boldsymbol{X}) \end{pmatrix}$$

where c is some positive constant. Show that $E(Y|\boldsymbol{X}) = \alpha + \boldsymbol{\beta}^T\boldsymbol{X}$ where

$$\alpha = \mu_Y - \boldsymbol{\beta}^T\boldsymbol{\mu}_X \quad \text{and}$$

$$\boldsymbol{\beta} = [\mathrm{Cov}(\boldsymbol{X})]^{-1}\mathrm{Cov}(\boldsymbol{X}, Y).$$

3.11. (Due to R.D. Cook.) Let $\boldsymbol{X}$ be a $p \times 1$ random vector with $E(\boldsymbol{X}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$. Let $\boldsymbol{B}$ be any constant full rank $p \times r$ matrix where $1 \leq r \leq p$. Suppose that for all such conforming matrices $\boldsymbol{B}$,

$$E(\boldsymbol{X}|\boldsymbol{B}^T\ \boldsymbol{X}) = \boldsymbol{M}_B\boldsymbol{B}^T\boldsymbol{X}$$

where $\boldsymbol{M}_B$ a $p \times r$ constant matrix that depend on $\boldsymbol{B}$.

Using the fact that $\boldsymbol{\Sigma B} = \mathrm{Cov}(\boldsymbol{X}, \boldsymbol{B}^{\mathrm{T}}\boldsymbol{X}) = \mathrm{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{B}) = \mathrm{E}[\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{B}|\boldsymbol{B}^{\mathrm{T}}\boldsymbol{X})]$, compute $\boldsymbol{\Sigma B}$ and show that $\boldsymbol{M}_B = \boldsymbol{\Sigma B}(\boldsymbol{B}^T\boldsymbol{\Sigma B})^{-1}$. Hint: what acts as a constant in the inner expectation?

3.12. Let $\boldsymbol{x}$ be a $p \times 1$ random vector with covariance matrix $\mathrm{Cov}(\boldsymbol{x})$. Let $\boldsymbol{A}$ be an $r \times p$ constant matrix and let $\boldsymbol{B}$ be a $q \times p$ constant matrix. Find $\mathrm{Cov}(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{B}\boldsymbol{x})$ in terms of $\boldsymbol{A}, \boldsymbol{B}$, and $\mathrm{Cov}(\boldsymbol{x})$.

3.13. The table $\boldsymbol{W}$ shown below represents 4 measurements on 5 people.

age	breadth	cephalic	size
39.00	149.5	81.9	3738
35.00	152.5	75.9	4261
35.00	145.5	75.4	3777
19.00	146.0	78.1	3904
0.06	88.5	77.6	933

a) Find the sample mean $\overline{\boldsymbol{x}}$.
b) Find the coordinatewise median MED($\boldsymbol{W}$).

3.14. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid $p \times 1$ random vectors from a multivariate t-distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ with d degrees of freedom. Then $E(\boldsymbol{x}_i) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{x}) = \dfrac{d}{d-2}\boldsymbol{\Sigma}$ for $d > 2$. Assuming $d > 2$, find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

3.15. Suppose that

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim N_4 \left(\begin{pmatrix} 9 \\ 16 \\ 4 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 & -0.4 & 0 \\ 0.8 & 1 & -0.56 & 0 \\ -0.4 & -0.56 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \right).$$

a) Find the distribution of X_3.
b) Find the distribution of $(X_2, X_4)^T$.
c) Which pairs of random variables X_i and X_j are independent?
d) Find the correlation $\rho(X_1, X_3)$.

3.16. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid $p \times 1$ random vectors where

$$\boldsymbol{x}_i \sim (1-\gamma)N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \gamma N_p(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$$

with $0 < \gamma < 1$ and $c > 0$. Then $E(\boldsymbol{x}_i) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{x}_i) = [1 + \gamma(c-1)]\boldsymbol{\Sigma}$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{d})$ for appropriate vector $\boldsymbol{d}$.

3.17. Let $\boldsymbol{X}$ be an $n \times p$ constant matrix and let $\boldsymbol{\beta}$ be a $p \times 1$ constant vector. Suppose $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$. Find the distribution of $\boldsymbol{H}\boldsymbol{Y}$ if $\boldsymbol{H}^T = \boldsymbol{H} = \boldsymbol{H}^2$ is an $n \times n$ matrix and if $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$. Simplify.

3.18. Recall that if $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal with mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. Let Y and X follow a bivariate normal distribution

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2\left(\begin{pmatrix} 134 \\ 96 \end{pmatrix}, \begin{pmatrix} 24.5 & 1.1 \\ 1.1 & 23.0 \end{pmatrix} \right).$$

a) Find $E(Y|X)$.

b) Find $\mathrm{Var}(Y|X)$.

3.19. Suppose that

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim N_4\left(\begin{pmatrix} 1 \\ 7 \\ 3 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & 0 & 2 & 1 \\ 0 & 1 & 0 & 0 \\ 2 & 0 & 3 & 1 \\ 1 & 0 & 1 & 5 \end{pmatrix} \right).$$

a) Find the distribution of $(X_1, X_4)^T$.

b) Which pairs of random variables X_i and X_j are independent?

c) Find the correlation $\rho(X_1, X_4)$.

3.20. Suppose that

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim N_4\left(\begin{pmatrix} 3 \\ 4 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 3 & 2 & 1 & 1 \\ 2 & 4 & 1 & 0 \\ 1 & 1 & 2 & 0 \\ 1 & 0 & 0 & 3 \end{pmatrix} \right).$$

a) Find the distribution of $(X_1, X_3)^T$.

b) Which pairs of random variables X_i and X_j are independent?

c) Find the correlation $\rho(X_1, X_3)$.

3.21. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid $p \times 1$ random vectors where $E(\boldsymbol{x}_i) = e^{0.5}\mathbf{1}$ and $\mathrm{Cov}(\boldsymbol{x}_i) = (e^2 - e)\boldsymbol{I}_p$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

3.22. Suppose that

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim N_4\left(\begin{pmatrix} 49 \\ 25 \\ 9 \\ 4 \end{pmatrix}, \begin{pmatrix} 2 & -1 & 3 & 0 \\ -1 & 5 & -3 & 0 \\ 3 & -3 & 5 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix} \right).$$

a) Find the distribution of $(X_1, X_3)^T$.

b) Which pairs of random variables X_i and X_j are independent?

c) Find the correlation $\rho(X_1, X_3)$.

3.23. Recall that if $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal with mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. Let Y and X follow a bivariate normal distribution

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2\left(\begin{pmatrix} 49 \\ 17 \end{pmatrix}, \begin{pmatrix} 3 & -1 \\ -1 & 4 \end{pmatrix} \right).$$

a) Find $E(Y|X)$.

b) Find $\mathrm{Var}(Y|X)$.

3.24. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid 2×1 random vectors from a multivariate lognormal LN($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$) distribution. Let $\boldsymbol{x}_i = (X_{i1}, X_{i2})^T$. Following Press (2005, pp. 149-150), $E(X_{ij}) = \exp(\mu_j + \sigma_j^2/2)$,
$V(X_{ij}) = \exp(\sigma_j^2)[\exp(\sigma_j^2) - 1]\exp(2\mu_j)$ for $j = 1, 2$, and
$\mathrm{Cov}(X_{i1}, X_{i2}) = \exp[\mu_1 + \mu_2 + 0.5(\sigma_1^2 + \sigma_2^2) + \sigma_{12}][\exp(\sigma_{12}) - 1]$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

3.25. Following Srivastava and Khatri (1979, p. 47), let

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} \sim N_p\left[\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right].$$

a) Show that the nonsingular linear transformation

$$\begin{pmatrix} \boldsymbol{I} & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2 \\ \boldsymbol{X}_2 \end{pmatrix} \sim$$

$$N_p\left[\begin{pmatrix} \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right].$$

b) Then $\boldsymbol{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2 \perp\!\!\!\perp \boldsymbol{X}_2$, and

$$\boldsymbol{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2 \sim N_q(\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

By independence, $\boldsymbol{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2$ has the same distribution as $(\boldsymbol{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2)|\boldsymbol{X}_2$, and the term $-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2$ is a constant, given $\boldsymbol{X}_2$. Use this result to show that

$$\boldsymbol{X}_1|\boldsymbol{X}_2 \sim N_q(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{X}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

R Problems Use the command *source("G:/rpack.txt")* **to download the functions** and the command *source("G:/robdata.txt")* **to download the data. See Preface or Section 11.2.** Typing the name of the `rpack` function, e.g. *covmba*, will display the code for the function. Use the `args` command, e.g. *args(covmba)*, to display the needed arguments for the function. For some of the following problems, the *R* commands can be copied and pasted from (http://parker.ad.siu.edu/Olive/robRhw.txt) into *R*.

3.26. a) Download the `maha` function that creates the classical Mahalanobis distances.

b) Copy and paste the commands for this problem and check whether observations 1–40 look like outliers.

3.27. Download the `rmaha` function that creates the robust Mahalanobis distances using `cov.mcd` (FMCD). Obtain outx2 as in Problem 3.26 b). Enter the *R* command *library(MASS)*. Enter the command *rmaha(outx2)* and check whether observations 1–40 look like outliers.

3.28. a) Download the `covmba` function.

b) Download the program `rcovsim`.

c) Enter the command `rcovsim(100)` three times and include the output in *Word*.

d) Explain what the output is showing.

3.29*. a) Assuming that you have done the two source commands above Problem 3.26 (and the *R* command *library(MASS)*), type the command *ddcomp(buxx)*. This will make 4 DD plots based on the DGK, FCH, FMCD, and median ball estimators. The DGK and median ball estimators are the two attractors used by the FCH estimator. With the leftmost mouse button, move the cursor to an outlier and click. This data is the Buxton (1920) data and cases with numbers 61, 62, 63, 64, and 65 were the outliers with head lengths near 5 feet. After identifying at least three outliers in each plot, hold the rightmost mouse button down (and in *R* click on *Stop*) to advance to the next plot. When done, hold down the *Ctrl* and *c* keys to make a copy of the plot. Then paste the plot in *Word*.

b) Repeat a) but use the command *ddcomp(cbrainx)*. This data is the Gladstone (1905) data and some infants are multivariate outliers.

c) Repeat a) but use the command *ddcomp(museum[,-1])*. This data is the Schaaffhausen (1878) skull measurements and cases 48–60 were apes while the first 47 cases were humans.

3.30*. (Perform the *source("G:/rpack.txt")* command if you have not already done so.) The *concmv* function illustrates concentration with $p = 2$ and a scatterplot of X_1 versus X_2. The outliers are such that the MBA and FCH estimators can not always detect them. Type the command *concmv()*. Hold

the rightmost mouse button down (and in *R* click on *Stop*) to see the DD plot after one concentration step. The start uses the coordinatewise median and $diag([MAD(X_i)]^2)$. Repeat 4 more times to see the DD plot based on the attractor. The outliers have large values of X_2 and the highlighted cases have the smallest distances. Repeat the command *concmv()* several times. Sometimes the start will contain outliers but the attractor will be clean (none of the highlighted cases will be outliers), but sometimes concentration causes more and more of the highlighted cases to be outliers, so that the attractor is worse than the start. Copy one of the DD plots where none of the outliers are highlighted into *Word*.

3.31***.** (Perform the *source("G:/rpack.txt")* command if you have not already done so.) The *ddmv* function illustrates concentration with the DD plot. The outliers are highlighted. The first graph is the DD plot after one concentration step. Hold the rightmost mouse button down (and in *R* click on *Stop*) to see the DD plot after two concentration steps. Repeat 4 more times to see the DD plot based on the attractor. In this problem, try to determine the proportion of outliers *gam* that the DGK estimator can detect for $p = 2, 4, 10$ and 20. Make a table of p and *gam*. For example the command *ddmv(p=2,gam=.4)* suggests that the DGK estimator can tolerate nearly 40% outliers with $p = 2$, but the command *ddmv(p=4,gam=.4)* suggest that *gam* needs to be lowered (perhaps by 0.1 or 0.05). Try to make $0 < gam < 0.5$ as large as possible.

3.32. (Perform the *source("G:/rpack.txt")* command if you have not already done so.) A simple modification of the MBA estimator adds starts trimming M% of cases furthest from the coordinatewise median MED($\boldsymbol{x}$). For example use $M \in \{98, 95, 90, 80, 70, 60, 50\}$. Obtain the program *cmba2* from `rpack.txt` and try the MBA estimator on the data sets in Problem 3.29.

3.33. The *rpack* function `covesim` compares various ways to robustly estimate the covariance matrix. The estimators used are ccov: the classical estimator applied to the clean cases, RFCH, and RMVN. The average dispersion matrix is reported over nruns = 20. Let diag(A) be the diagonal of the average dispersion matrix. Then diagdiff = diag(ccov) - diag(rmvne) and abssumd = sum(abs(diagdiff)). The clean data $\sim N_p(0, diag(1, ..., p))$.

a) The *R* command covesim(n=100,p=4) gives output when there are no outliers. Copy and paste the output into *Word*.

b) The command covesim(n=100,p=4,outliers=1,pm=15) uses 40% outliers that are a tight cluster at major axis with mean $(0, ..., 0, pm)^T$. Hence *pm* determines how far the outliers are from the bulk of the data. Copy and paste the output into *Word*. The average dispersion matrices should be $\approx c\ diag(1, 2, 3, 4)$ for this type of outlier configuration. What is c for RFCH and RMVN?

3.34. The *R* function `cov.mcd` is an FMCD estimator. If `cov.mcd` computed the minimum covariance determinant estimator, then the log determinant of the dispersion matrix would be a minimum and would not change when the rows of the data matrix are permuted. The *R commands* for this problem permute the rows of the Gladstone (1905) data matrix seven times. The log determinant is given for each of the resulting `cov.mcd` estimators.

a) Paste the output into *Word*.

b) How many distinct values of the log determinant were produced? (Only one if the MCD estimator is being computed.)

3.35. 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 10 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 $p = 3$.)

3.36. 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 $p = 3$.)

3.37*. 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?

3.38. 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)
```

3.39. a) Download the function `mplot`.

b) Enter the commands in Problem 3.37b 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.

3.40 a) Download the function `wddplot`.

b) Enter the commands in Problem 3.37b to obtain a data set x.

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

3.41. 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 region uses the $100(1 - \delta + p/n) = 84.6$th percentile. 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. (Right click *Stop* once on the plot.)

3.42. For the Buxton (1920) data with multiple linear regression, *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!

a) Copy and paste the commands for this problem into *R*. Include the lasso response plot in *Word*. The identity line passes right through the outliers which are obvious because of the large gap. Prediction interval (PI) bands are also included in the plot.

b) Copy and paste the commands for this problem into *R*. Include the lasso response plot in *Word*. This did lasso for the cases in the `covmb2` set B applied to the predictors which included all of the clean cases and omitted

the 5 outliers. The response plot was made for all of the data, including the outliers.

c) Copy and paste the commands for this problem into *R*. Include the DD plot in *Word*. The outliers are in the upper right corner of the plot.

3.43. The *rpack* function `mldsim6` compares 7 estimators: FCH, RFCH, CMVE, RCMVE, RMVN, `covmb2`, and MB described in Olive (2017b, ch. 4). Most of these estimators need $n > 2p$, need a nonsingular dispersion matrix, and work best with $n > 10p$. The function generates data sets and counts how many times the minimum Mahalanobis distance $D_i(T, \boldsymbol{C})$ of the outliers is larger than the maximum distance of the clean data. The value pm controls how far the outliers need to be from the bulk of the data, and pm roughly needs to increase with $\sqrt{p}$.

For data sets with $p > n$ possible, the function `mldsim7` used the Euclidean distances $D_i(T, \boldsymbol{I}_p)$ and the Mahalanobis distances $D_i(T, \boldsymbol{C}_d)$ where $\boldsymbol{C}_d$ is the diagonal matrix with the same diagonal entries as $\boldsymbol{C}$ where $(T, \boldsymbol{C})$ is the `covmb2` estimator using j concentration type steps. Dispersion matrices are effected more by outliers than good robust location estimators, so when the outlier proportion is high, it is expected that the Euclidean distances $D_i(T, \boldsymbol{I}_p)$ will outperform the Mahalanobis distance $D_i(T, \boldsymbol{C}_d)$ for many outlier configurations. Again the function counts the number of times the minimum outlier distance is larger than the maximum distance of the clean data.

Both functions used several outlier types. The simulations generated 100 data sets. The clean data had $\boldsymbol{x}_i \sim N_p(\boldsymbol{0}, diag(1, ..., p))$. Type 1 had outliers in a tight cluster (near point mass) at the major axis $(0, ..., 0, pm)^T$. Type 2 had outliers in a tight cluster at the minor axis $(pm, 0, ..., 0)^T$. Type 3 had mean shift outliers $\boldsymbol{x}_i \sim N_p((pm, ..., pm)^T, diag(1, ..., p))$. Type 4 changed the pth coordinate of the outliers to pm. Type 5 changed the 1st coordinate of the outliers to pm. (If the outlier $\boldsymbol{x}_i = (x_{1i}, ..., x_{pi})^T$, then $x_{i1} = pm$.)

Table 3.7 Number of Times All Outlier Distances > Clean Distances, otype=1

n	p	γ	osteps	pm	FCH	RFCH	CMVE	RCMVE	RMVN	covmb2	MB
100	10	0.25	0	20	85	85	85	85	86	67	89

a) Table 3.7 suggests with osteps = 0, `covmb2` had the worst count. When pm is increased to 25, all counts become 100. Copy and paste the commands for this part into *R* and make a table similar to Table 3.7, but now osteps=9 and $p = 45$ is close to $n/2$ for the second line where $pm = 60$. Your table should have 2 lines from output.

b) Copy and paste the commands for this part into *R* and make a table similar to Table 3.8, but type 2 outliers are used.

c) When you have two reasonable outlier detectors, there are outlier configurations where one will beat the other. Simulations suggest that "covmb2"

Table 3.8 Number of Times All Outlier Distances > Clean Distances, otype=1

n	p	γ	osteps	pm	covmb2	diag
100	1000	0.4	0	1000	100	41
100	1000	0.4	9	600	100	42

using $D_i(T, \boldsymbol{I}_p)$ outperforms "diag" using $D_i(T, \boldsymbol{C}_d)$ for many outlier configurations, but there are some exceptions. Copy and paste the commands for this part into R and make a table similar to Table 3.8, but type 3 outliers are used.

3.44. Tests for covariance matrices tend to be 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 $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{\Sigma}_0$ for known $\boldsymbol{\Sigma}_0$ is to plot $D_i(\overline{\boldsymbol{x}}, \boldsymbol{S})$ versus $D_i(\overline{\boldsymbol{x}}, \boldsymbol{\Sigma}_0)$ for $i = 1, ..., n$. If $n \geq 10p$ and H_0 is true, then the plotted points in the DD plot should start to cluster tightly about the identity line.

a) A test for sphericity is a test of $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}} = \sigma^2 \boldsymbol{I}_p$ for some unknown constant $\sigma^2 > 0$. Make a "D^2 plot" of $D_i^2(\overline{\boldsymbol{x}}, \boldsymbol{S})$ versus $D_i^2(\overline{\boldsymbol{x}}, \boldsymbol{I}_p)$. If $n \geq 10p$ and H_0 is true, then the plotted points in the D^2 plot should cluster tightly about the line through the origin with slope σ^2. 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 we 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, consider 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 $n_i \geq 10p$ and H_0 is true, what line will the plotted points cluster about in each of the k DD plots? (See Equation (8.2) for $\boldsymbol{S}_{pool}$.)

Remark 3.11. Lots of other diagnostic DD plots can be made. Suppose known parts of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ are hypothesized to be $\boldsymbol{0}$. Let $\boldsymbol{S}_Z$ be the sample covariance matrix with the known parts set to $\boldsymbol{0}$. Then plot $D_i(\overline{\boldsymbol{x}}, \boldsymbol{S})$ versus $D_i(\overline{\boldsymbol{x}}, \boldsymbol{S}_Z)$. For example, a diagnostic for $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}} = diag(\boldsymbol{\Sigma}_{11}, ..., \boldsymbol{\Sigma}_{kk})$ where the $\boldsymbol{\Sigma}_{ii}$ are unknown block matrices is the above plot with $\boldsymbol{S}_Z = diag(\boldsymbol{S}_{11}, ..., \boldsymbol{S}_{kk})$. A diagnostic for $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}} = diag(\sigma_{11}, ..., \sigma_{pp})$ where the σ_{ii} are unknown would use $\boldsymbol{S}_Z = diag(s_{11}, ..., s_{pp})$ if $\boldsymbol{S} = (s_{ij})$. Another diagnostic would check whether the population correlation matrix $\boldsymbol{\rho}_{\boldsymbol{x}} = \boldsymbol{I}_p$. See the following paragraph.

Similar diagnostic DD plots can be made for the population correlation matrix $\boldsymbol{\rho}_{\boldsymbol{x}}$ where scaled data $\boldsymbol{z}_i$ is used in the D_i such that the sample mean of the scaled data is $\overline{\boldsymbol{z}} = \boldsymbol{0}$ and the sample covariance matrix of the scaled data is $\boldsymbol{S}_{\boldsymbol{z}} = \boldsymbol{R} = (r_{ij})$. If the data matrix is x with rows $\boldsymbol{x}_i^T$, then the R command

```
z <- scale(x)
```

will make a data matrix z with rows $\boldsymbol{z}_i^T$. For example, consider $H_0 : \boldsymbol{\rho_x} = \boldsymbol{\rho}_0 = (\rho_{ij})$ where $\rho_{ij} = \rho$ for $i \neq j$ where $-1 < \rho < 1$ is unknown, and $\rho_{ii} = 1$ for $i = 1, ..., p$. Let $\hat{\rho}$ be the average of the r_{ij} where $i < j$. Let $\boldsymbol{R}_r = (p_{ij})$ where $p_{ij} = \hat{\rho}$ for $i \neq j$ and $p_{ii} = 1$ for $i = 1, ..., p$. Then make a DD plot of $D_i(\mathbf{0}, \boldsymbol{R})$ versus $D_i(\mathbf{0}, \boldsymbol{R}_r)$.

The RMVN matrix $\boldsymbol{C}_{RMVN}$ could be used in place of $\boldsymbol{S}$ in some of the plots if $\boldsymbol{C}_{RMVN} \xrightarrow{P} c\boldsymbol{\Sigma_x}$ for some constant $c > 0$. Then for some of the plots the plotted points might scatter about some line through the origin instead of the identity line.