# Chapter 10

# Multivariate Models

**Definition 10.1.** An important *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 $\boldsymbol{x}$ that is completely specified by a $p \times 1$ population *location* vector $\boldsymbol{\mu}$ and a $p \times p$ symmetric positive definite population *dispersion* matrix $\boldsymbol{\Sigma}$. Thus $P(\boldsymbol{x} \in A) = \int_A f(\boldsymbol{z}) d\boldsymbol{z}$ for suitable sets $A$.

The multivariate location and dispersion model is in many ways similar to the multiple linear regression model. 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.

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

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

where the $i$th row of $\boldsymbol{W}$ is $\boldsymbol{x}_i^T$ and the $j$th column is $\boldsymbol{w}^j$. Each column $\boldsymbol{w}^j$ of $\boldsymbol{W}$ corresponds to a variable. 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.

There are some differences in the notation used in multiple linear regression and multivariate location dispersion models. Notice that $\boldsymbol{W}$ could be used as the design matrix in multiple linear regression although usually the first column of the regression design matrix is a vector of ones. The $n \times p$ design matrix in the multiple linear regression model was denoted by $\boldsymbol{X}$ and $X_i \equiv \boldsymbol{x}^i$ denoted the $i$th column of $\boldsymbol{X}$. In the multivariate location dispersion model, $\boldsymbol{X}$ and $\boldsymbol{X}_i$ will be used to denote a $p \times 1$ random vector with observed value $\boldsymbol{x}_i$, and $\boldsymbol{x}_i^T$ is the $i$th row of the data matrix $\boldsymbol{W}$. Johnson and Wichern (1988, p. 7, 53) uses $\boldsymbol{X}$ to denote the $n \times p$ data matrix and a $n \times 1$ 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/Splus* 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 $x$.

## 10.1 The Multivariate Normal Distribution

**Definition 10.2: 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

$$f(\boldsymbol{z}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(1/2)(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})} \tag{10.1}$$

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 10.3.** 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*

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

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

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

$$E(\boldsymbol{a} + \boldsymbol{X}) = \boldsymbol{a} + E(\boldsymbol{X}) \text{ and } E(\boldsymbol{X} + \boldsymbol{Y}) = E(\boldsymbol{X}) + E(\boldsymbol{Y}) \tag{10.2}$$

and

$$E(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}E(\boldsymbol{X}) \text{ and } E(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}) = \boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}. \tag{10.3}$$

Thus

$$\text{Cov}(\boldsymbol{a} + \boldsymbol{A}\boldsymbol{X}) = \text{Cov}(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}\text{Cov}(\boldsymbol{X})\boldsymbol{A}^T. \tag{10.4}$$

Some important properties of MVN distributions are given in the following three propositions. These propositions can be proved using results from Johnson and Wichern (1988, p. 127-132).

**Proposition 10.1.** a) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and

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

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 $\sigma_{i,j} = 0$ while the diagonal entries of $\boldsymbol{\Sigma}$ are $\sigma_{i,i} = \sigma_i^2$).

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, then $\boldsymbol{a} + \boldsymbol{X} \sim N_p(\boldsymbol{a} + \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

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

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

**Proposition 10.2.** 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

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

**Proposition 10.3. 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 $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. 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 10.1.** 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 recall that 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 10.1.** 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 Proposition 10.1b and Proposition 10.2c 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. See Seber and Lee (2003, p. 23), Kowalski (1973) and 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 $f(x,y) =$

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

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

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 Proposition 10.2 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 10.2.** In Proposition 10.3, suppose that $\boldsymbol{X} = (Y, X_2, ..., X_p)^T$. Let $X_1 = Y$ and $\boldsymbol{X}_2 = (X_2, ..., X_p)^T$. Then $E[Y|\boldsymbol{X}_2] = \beta_1 + \beta_2X_2 + \cdots + \beta_pX_p$ and $\text{VAR}[Y|\boldsymbol{X}_2]$ is a constant that does not depend on $\boldsymbol{X}_2$. Hence $Y = \beta_1 + \beta_2X_2 + \cdots + \beta_pX_p + e$ follows the multiple linear regression model.

## 10.2 Elliptically Contoured Distributions

**Definition 10.4: 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 density

$$f(\boldsymbol{z}) = k_p |\boldsymbol{\Sigma}|^{-1/2} g[(\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu})], \tag{10.5}$$

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

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp(i\boldsymbol{t}^T\boldsymbol{\mu})\psi(\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}) \tag{10.6}$$

for some function $\psi$. If the second moments exist, then

$$E(\boldsymbol{X}) = \boldsymbol{\mu} \tag{10.7}$$

and

$$\mathrm{Cov}(\boldsymbol{X}) = c_X \boldsymbol{\Sigma} \tag{10.8}$$

where

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

**Definition 10.5.** The *population squared Mahalanobis distance*

$$U \equiv D^2 = D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \tag{10.9}$$

has density

$$h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)} k_p u^{p/2-1} g(u). \tag{10.10}$$

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 $\chi^2_p$ density.

The following lemma is useful for proving properties of EC distributions without using the characteristic function (10.6). See Eaton (1986) and Cook (1998a, p. 57, 130).

**Lemma 10.4.** Let $\boldsymbol{X}$ be a $p \times 1$ random vector with 1st moments; ie, $E(\boldsymbol{X})$ 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 $\boldsymbol{B}$,

$$E(\boldsymbol{X}|\boldsymbol{B}^T\ \boldsymbol{X}) = \boldsymbol{\mu} + \boldsymbol{M}_B\boldsymbol{B}^T(\boldsymbol{X} - \boldsymbol{\mu}) = \boldsymbol{a}_B + \boldsymbol{M}_B\boldsymbol{B}^T\boldsymbol{X} \tag{10.11}$$

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

To use this lemma to prove interesting properties, 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

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

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

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

Another 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 10.11. 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 2nd moments, $\text{Cov}(\boldsymbol{X})$ can be used instead of $\boldsymbol{\Sigma}$.

**Proposition 10.5.** 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 $\text{Cov}(\boldsymbol{X})$ is nonsingular,

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

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

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

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

by Lemma 10.4. 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 Lemma 10.4. Notice that $\boldsymbol{M}_B = \boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B})^{-1} =$

$$\begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{0} \end{pmatrix} [\begin{pmatrix} \boldsymbol{A}^T & \boldsymbol{0}^T \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{0} \end{pmatrix}]^{-1}$$

$$= \begin{pmatrix} \boldsymbol{M}_{1B} \\ \boldsymbol{M}_{2B} \end{pmatrix}.$$

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 $\boldsymbol{\Sigma}_{11}$. QED

**Proposition 10.6.** 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 $\alpha$ and $\boldsymbol{\beta}$ are given in a).

**Proof.** a) The trick is to choose $\boldsymbol{B}$ so that Lemma 10.4 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

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

Now

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

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

by Lemma 10.4. 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, Dehon, Rousseeuw and Van Aelst (2001) for references.

**Example 10.2.** This example illustrates another application of Lemma 10.4. 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 Proposition 4.1c),

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

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

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 Lemma 10.4.

## 10.3 Sample Mahalanobis Distances

In the multivariate location and dispersion model, sample Mahalanobis distances play a role similar to that of residuals in multiple linear regression. The observed data $\boldsymbol{X}_i = \boldsymbol{x}_i$ for $i = 1, ..., n$ is 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 covariance estimator.

**Definition 10.6.** The $i$th *squared Mahalanobis distance* is

$$D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{x}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T(\boldsymbol{W})) \quad (10.12)$$

for each point $\boldsymbol{x}_i$. Notice that $D_i^2$ is a random variable (scalar valued).

Notice that the population squared Mahalanobis distance is

$$D_{\boldsymbol{x}}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \quad (10.13)$$

and 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 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.

**Example 10.3.** 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 $R_\alpha$ 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

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

where $P(W \leq \chi_p^2(\alpha)) = \alpha$ if $W \sim \chi_p^2$. 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 $\alpha$-density region will use a constant $U_\alpha$ obtained from Equation (10.10).

The classical Mahalanobis distance corresponds to the sample mean and sample covariance matrix

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

and

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

and will be denoted by $MD_i$. 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 $RD_i$.

## 10.4 Affine Equivariance

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 $\boldsymbol{B} = \mathbf{1}\boldsymbol{b}^T$ where $\mathbf{1}$ is an $n \times 1$ vector of ones and $\boldsymbol{b}$ is a $p \times 1$ constant vector. Hence the $i$th 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 $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{A} + \boldsymbol{B}$ where $\boldsymbol{A}$ is any nonsingular $p \times p$ matrix.

**Definition 10.7.** Then the multivariate location and dispersion estimator $(T, \boldsymbol{C})$ is *affine equivariant* if

$$T(\boldsymbol{Z}) = T(\boldsymbol{W}\boldsymbol{A} + \boldsymbol{B}) = \boldsymbol{A}^T T(\boldsymbol{W}) + \boldsymbol{b}, \tag{10.14}$$

and

$$\boldsymbol{C}(\boldsymbol{Z}) = \boldsymbol{C}(\boldsymbol{W}\boldsymbol{A} + \boldsymbol{B}) = \boldsymbol{A}^T\boldsymbol{C}(\boldsymbol{W})\boldsymbol{A}. \tag{10.15}$$

The following proposition shows that the Mahalanobis distances are invariant under affine transformations. See Rousseeuw and Leroy (1987, p. 252-262) for similar results.

**Proposition 10.7.** If $(T, \boldsymbol{C})$ is affine equivariant, then

$$D_i^2(\boldsymbol{W}) \equiv D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) =$$

$$D_i^2(T(\boldsymbol{Z}), \boldsymbol{C}(\boldsymbol{Z})) \equiv D_i^2(\boldsymbol{Z}). \tag{10.16}$$

**Proof.** Since $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{A} + \boldsymbol{B}$ has $i$th row

$$\boldsymbol{z}_i^T = \boldsymbol{x}_i^T\boldsymbol{A} + \boldsymbol{b}^T,$$

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

## 10.5 Breakdown

This section gives a standard definition of breakdown for estimators of multivariate location and dispersion. The following notation will be useful. Let $\boldsymbol{W}$ denote the $n \times p$ data matrix with $i$th row $\boldsymbol{x}_i^T$ corresponding to the $i$th case. Let $\boldsymbol{W}_d^n$ denote the data matrix with $i$th row $\boldsymbol{w}_i^T$ where any $d$ of the cases have been replaced by arbitrarily bad contaminated cases. Then the contamination fraction is $\gamma = d/n$. 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.

**Definition 10.8.** The breakdown value of the multivariate location estimator $T$ at $\boldsymbol{W}$ is

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

where the supremum is over all possible corrupted samples $\boldsymbol{W}_d^n$ and $1 \leq d \leq n$. Let $0 \leq \lambda_p(\boldsymbol{C}(\boldsymbol{W})) \leq \cdots \leq \lambda_1(\boldsymbol{C}(\boldsymbol{W}))$ 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 $\infty$. Hence the breakdown value of the dispersion estimator is

$$B(\boldsymbol{C}, \boldsymbol{W}) = \min\{\frac{d}{n} : \sup_{\boldsymbol{W}_d^n} \text{med}[\frac{1}{\lambda_p(\boldsymbol{C}(\boldsymbol{W}_d^n))}, \lambda_1(\boldsymbol{C}(\boldsymbol{W}_d^n))] = \infty\}.$$

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 $i$th 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.

**Proposition 10.8.** If nonequivariant estimators (that 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 $d_T$ is the smallest number of arbitrarily bad cases that can make the median Euclidean distance $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_{d_T}^n)\|)$ arbitrarily large.

**Proof.** Note that for a fixed data set $\boldsymbol{W}_d^n$ with $i$th row $\boldsymbol{w}_i$, if the multivariate location estimator $T(\boldsymbol{W}_d^n)$ satisfies $\|T(\boldsymbol{W}_d^n)\| = M$ for some constant $M$, then the median Euclidean distance $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|) \leq \max_{i=1,...,n} \|\boldsymbol{x}_i - T(\boldsymbol{W}_d^n)\| \leq \max_{i=1,...,n} \|\boldsymbol{x}_i\| + M$ if $d < n/2$. Similarly, if $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|) = M$ for some constant $M$, then $\|T(\boldsymbol{W}_d^n)\|$ is bounded if $d < n/2$. QED

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. The following result 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.

**Proposition 10.9.** 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.

**Proof.** Note that $\overline{\boldsymbol{x}}_J$ is HB by Proposition 10.8. From numerical linear algebra, it is known that the largest eigenvalue of a $p \times p$ matrix $\boldsymbol{C}$ is bounded above by $p \max |c_{i,j}|$ where $c_{i,j}$ is the $(i,j)$ entry of $\boldsymbol{C}$. See Datta (1995, p. 403). Denote the $c_n$ cases by $\boldsymbol{z}_1, ..., \boldsymbol{z}_{c_n}$. Then the $(i,j)$th element $c_{i,j}$ of $\boldsymbol{C} \equiv \boldsymbol{S}_J$ is

$$c_{i,j} = \frac{1}{c_n - 1} \sum_{k=1}^{c_n} (z_{i,k} - \overline{z}_k)(z_{j,k} - \overline{z}_j).$$

Hence the maximum eigenvalue $\lambda_1$ is bounded if fewer than half of the cases are outliers. Unless the percentage of outliers is high (higher than a value tending to 0.5 as $n \to \infty$), the determinant $|\boldsymbol{C}_{MCD}(c_n)|$ of the HB minimum covariance determinant (MCD) estimator of Definition 10.9 below is greater than 0. Thus $0 < |\boldsymbol{C}_{MCD}(c_n)| \leq |\boldsymbol{S}_J| = \lambda_1 \cdots \lambda_p$, and $\lambda_p > |\boldsymbol{C}_{MCD}(c_n)|/\lambda_1^{p-1} > 0$. QED

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

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq D^2_{(c_n)}\} \tag{10.17}$$

where $D^2_{(c_n)}$ is the $c_n$th smallest squared Mahalanobis distance based on $(T, \boldsymbol{C})$. This ellipsoid contains the $c_n$ cases with the smallest $D^2_i$. The volume of this ellipsoid is proportional to the square root of the determinant $|\boldsymbol{C}|^{1/2}$, and this volume will be positive unless extreme degeneracy is present among the $c_n$ cases. See Johnson and Wichern (1988, p. 103-104).

## 10.6 Algorithms for the MCD Estimator

**Definition 10.9.** 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 $(T_{MCD}(\boldsymbol{W}), \boldsymbol{C}_{MCD}(\boldsymbol{W}))$.

The MCD estimator is a high breakdown 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. The population analog of the MCD estimator is closely related to the ellipsoid of highest concentration that contains $c_n/n \approx$ half of the mass. The MCD estimator is a $\sqrt{n}$ consistent HB estimator for

$$(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$$

where $a_{MCD}$ is some positive constant when the data $\boldsymbol{X}_i$ are elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$, and $T_{MCD}$ has a Gaussian limit. See Butler, Davies, and Jhun (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) note 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 high breakdown (HB) algorithms will again be used to approximate the robust estimators. Many of the properties and techniques used for HB regression algorithm estimators carry over for HB algorithm estimators of multivariate location and dispersion. Elemental sets are the key ingredient for both *basic resampling* and *concentration* algorithms.

**Definition 10.10.** 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 $(T_{0,j}, \boldsymbol{C}_{0,j})$ be the $j$th start (not necessarily elemental) and compute all $n$ Mahalanobis distances $D_i(T_{0,j}, \boldsymbol{C}_{0,j})$. At the next iteration, the classical estimator $(T_{1,j}, \boldsymbol{C}_{1,j}) = (\overline{\boldsymbol{x}}_{1,j}, \boldsymbol{S}_{1,j})$ 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 $(T_{0,j}, \boldsymbol{C}_{0,j}), (T_{1,j}, \boldsymbol{C}_{1,j}), ..., (T_{k,j}, \boldsymbol{C}_{k,j})$. The result of the iteration $(T_{k,j}, \boldsymbol{C}_{k,j})$ is called the $j$th attractor. If $K_n$ starts are used, then $j = 1, ..., K_n$. The concentration estimator $(T_{CMCD}, \boldsymbol{C}_{CMCD})$, called the CMCD estimator, is the attractor that has the smallest determinant $\det(\boldsymbol{C}_{k,j})$. The *basic resampling algorithm* estimator is a special case where $k = 0$ so that the attractor is the start: $(\overline{\boldsymbol{x}}_{k,j}, \boldsymbol{S}_{k,j}) = (\overline{\boldsymbol{x}}_{0,j}, \boldsymbol{S}_{0,j})$.

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.

**Proposition 10.10: Rousseeuw and Van Driessen (1999, p. 214).** Suppose that the classical estimator $(\overline{\boldsymbol{x}}_{i,j}, \boldsymbol{S}_{i,j})$ is computed from $c_n$ cases and that the $n$ Mahalanobis distances $\mathrm{RD}_m \equiv \mathrm{RD}_m(\overline{\boldsymbol{x}}_{i,j}, \boldsymbol{S}_{i,j})$ are computed. If $(\overline{\boldsymbol{x}}_{i+1,j}, \boldsymbol{S}_{i+1,j})$ is the classical estimator computed from the $c_n$ cases with the smallest Mahalanobis distances $\mathrm{RD}_m$, then the MCD criterion $\det(\boldsymbol{S}_{i+1,j}) \leq \det(\boldsymbol{S}_{i,j})$ with equality iff $(\overline{\boldsymbol{x}}_{i+1,j}, \boldsymbol{S}_{i+1,j}) = (\overline{\boldsymbol{x}}_{i,j}, \boldsymbol{S}_{i,j})$.

As in regression, starts that use a consistent initial estimator could be used. $K_n$ is the number starts and $k$ is the number of concentration steps used in the algorithm. Lopuhaä (1999) shows that if $(\overline{\boldsymbol{x}}_{1,1}, \boldsymbol{S}_{1,1})$ is the sample mean and covariance matrix applied to the cases with the smallest $c_n$ Mahalanobis distances based on the initial estimator $(T_{0,1}, \boldsymbol{C}_{0,1})$, then $(\overline{\boldsymbol{x}}_{1,1}, \boldsymbol{S}_{1,1})$ has the same rate of convergence as the initial estimator. Assume $k$ is fixed. If a start $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$, then the attractor is a

consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ where $a, s > 0$ are some constants. If the start is inconsistent, then so is the attractor. Hence the rate of the best attractor is equal to the rate of the best start.

**Proposition 10.11.** If $K$ and $k$ are fixed and free of $n$ (eg $K = 500$), then the elemental concentration algorithm estimator is inconsistent.

**Proof.** Following Hawkins and Olive (2002), the sample mean $\overline{\boldsymbol{x}}$ 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 $h_n \equiv p + 1$ cases is an inconsistent estimator, so the $K$ starts and the $K$ attractors are inconsistent by Lopuhaä (1999). The final estimator is an attractor and thus inconsistent.

If concentration is iterated to convergence so that $k$ is not fixed, then it has not been proven that the attractor is inconsistent if elemental starts are used. It is possible to produce consistent estimators if $K \equiv K_n$ is allowed to increase to $\infty$.

**Remark 10.3.** Let $\gamma_o$ be the highest percentage of large outliers that an elemental concentration algorithm can detect reliably. For many data sets,

$$\gamma_o \approx \min(\frac{n-c}{n}, 1 - [1 - (0.2)^{1/K}]^{1/h})100\% \tag{10.18}$$

if $n$ is large and $h = p + 1$.

The proof of this remark is exactly the same as the proof of Proposition 9.1 and Equation (10.18) agrees very well with the Rousseeuw and Van Driessen (1999) simulation performed on the hybrid FMCD algorithm that uses both concentration and partitioning. Section 10.7 will provide more theory for the CMCD algorithms and will show that there exists a useful class of data sets where the elemental concentration algorithm can tolerate up to 25% massive outliers.

## 10.7 Theory for CMCD Estimators

This section presents the FCH estimator to be used along with the classical and FMCD estimators. Recall from Definition 10.10 that a *concentration algorithm* uses $K_n$ *starts* $(T_{0,j}, \boldsymbol{C}_{0,j})$. Each start is refined with $k$ concentration

steps, resulting in $K_n$ *attractors* $(T_{k,j}, \boldsymbol{C}_{k,j})$, and the final estimator is the attractor that optimizes the criterion.

Concentration algorithms have been used by several authors, and the *basic resampling algorithm* is a special case with $k = 0$. Using $k = 10$ concentration steps works well, and iterating until convergence is usually fast. The DGK estimator (Devlin, Gnanadesikan and Kettenring 1975, 1981) defined below is one example. Gnanadesikan and Kettenring (1972, p. 94–95) provide a similar algorithm. The DGK estimator is affine equivariant since the classical estimator is affine equivariant and Mahalanobis distances are invariant under affine transformations by Proposition 10.7.

**Definition 10.11.** The DGK estimator $(\overline{\boldsymbol{x}}_{k,0}, \boldsymbol{S}_{k,0}) = (T_{DGK}, \boldsymbol{C}_{DGK})$ uses the classical estimator computed from all $n$ cases as the only start.

**Definition 10.12.** The median ball (MB) estimator $(\overline{\boldsymbol{x}}_{k,50}, \boldsymbol{S}_{k,50}) = (T_{MB}, \boldsymbol{C}_{MB})$ uses the classical estimator computed from the $c_n \approx n/2$ cases with $D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p) = \|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\| \leq \text{MED}(D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p))$ as a start. So the half set of cases $\boldsymbol{x}_i$ closest to the coordinatewise median MED$(\boldsymbol{W})$ in Euclidean distance is used. Let $(\overline{\boldsymbol{x}}_{-1,50}, \boldsymbol{S}_{-1,50}) = (\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$. Then the MB estimator is also the attractor of $(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$.

Some observations on breakdown from Section 10.5 will be useful for creating a simple robust estimator. If $d$ of the cases have been replaced by arbitrarily bad contaminated cases, then the contamination fraction is $\gamma = d/n$. Then the breakdown value of a multivariate location estimator is the smallest value of $\gamma$ needed to make $\|T\|$ arbitrarily large, and $T$ will not break down if $T$ can not be driven out of some ball of (possibly huge) radius $R$ about MED$(\boldsymbol{W})$. The breakdown value of a dispersion estimator $\boldsymbol{C}$ is the smallest value of $\gamma$ needed to drive the smallest eigenvalue to zero or the largest eigenvalue to $\infty$. Section 10.5 showed that if $(T, \boldsymbol{C})$ is the classical estimator $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ applied to some subset $J$ of $c_n \approx n/2$ cases of the data, then the maximum eigenvalue $\lambda_1$ can not get arbitrarily large if the $c_n$ cases are all contained in some ball of radius $R$ about the origin. Hence all of the $\lambda_i$ are bounded, and $\lambda_p$ can only be driven to zero if the determinant of $\boldsymbol{C}$ can be driven to zero. Using the above ideas suggests the following three robust estimators which use the same two starts.

**Definition 10.13.** Let the $M$th start $(T_{0,M}, \boldsymbol{C}_{0,M}) = (\overline{\boldsymbol{x}}_{0,M}, \boldsymbol{S}_{0,M})$ be the classical estimator applied after trimming the $M\%$ of cases furthest in Euclidean distance from the coordinatewise median MED($\boldsymbol{W}$) where $M \in \{0, 50\}$. Then concentration steps are performed resulting in the $M$th attractor $(T_{k,M}, \boldsymbol{C}_{k,M}) = (\overline{\boldsymbol{x}}_{k,M}, \boldsymbol{S}_{k,M})$. The $M = 0$ attractor is the DGK estimator and the $M = 50$ attractor is the MB estimator. The MBA estimator uses the attractor with the smallest determinant as does the FCH estimator if $\|\overline{\boldsymbol{x}}_{k,0} - \text{MED}(\boldsymbol{W})\| \leq \text{MED}(D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p))$. If the DGK location estimator has a greater Euclidean distance from MED($\boldsymbol{W}$) than half the data, then FCH uses the median ball attractor. Let $(T_A, \boldsymbol{C}_A)$ be the attractor used. Then the estimator $(T_F, \boldsymbol{C}_F)$ takes $T_F = T_A$ and

$$\boldsymbol{C}_F = \frac{\text{MED}(D_i^2(T_A, \boldsymbol{C}_A))}{\chi^2_{p,0.5}} \boldsymbol{C}_A \tag{10.19}$$

where $\chi^2_{p,0.5}$ is the 50th percentile of a chi–square distribution with $p$ degrees of freedom and F is the MBA or FCH estimator. CMVE is like FCH but the MVE criterion $[\text{MED}(D_i(\overline{\boldsymbol{x}}_{k,M}, \boldsymbol{S}_{k,M}))]^p \sqrt{det(\boldsymbol{S}_{k,M})}$ is used instead of the MCD criterion $det(\boldsymbol{S}_{k,M})$.

The following assumption and remark will be useful for examining the statistical properties of multivariate location and dispersion (MLD) estimators.

**Assumption (E1)**: Assume that $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ with nonsingular $\text{Cov}(\boldsymbol{X}) = c_X \boldsymbol{\Sigma}$ for some constant $c_X > 0$.

Then from Definition 10.5, the *population squared Mahalanobis distance*

$$U \equiv D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \tag{10.20}$$

has density

$$h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)} k_p u^{p/2-1} g(u), \tag{10.21}$$

and the 50% highest density region has the form of the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu}) \leq U_{0.5}\}$$

where $U_{0.5}$ is the median of the distribution of $U$. For example, if the $\boldsymbol{X}$ are MVN, then $U$ has the $\chi^2_p$ distribution.

**Remark 10.4.**

a) Butler, Davies and Jhun (1993): The MCD($c_n$) estimator is a $\sqrt{n}$ consistent HB estimator for $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ where the constant $a_{MCD} > 0$ depends on the EC distribution.

b) Lopuhaä (1999): If $(T, \boldsymbol{C})$ is a consistent estimator for $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate $n^{\delta}$ where the constants $s > 0$ and $\delta > 0$, then the classical estimator $(\overline{\boldsymbol{x}}_M, \boldsymbol{S}_M)$ computed after trimming the $M\%$ (where $0 < M < 100$) of cases with the largest distances $D_i(T, \boldsymbol{C})$ is a consistent estimator for $(\boldsymbol{\mu}, a_M\boldsymbol{\Sigma})$ with the same rate $n^{\delta}$ where $a_M > 0$ is some constant. Notice that applying the classical estimator to the $c_n \approx n/2$ cases with the smallest distances corresponds to $M = 50$.

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

d) Pratt (1959): Let $K$ be a fixed positive integer and let the constant $a > 0$. Suppose that $(T_1, \boldsymbol{C}_1), ..., (T_K, \boldsymbol{C}_K)$ are $K$ consistent estimators of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ each with the same rate $n^{\delta}$. 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 $n^{\delta}$.

e) Olive (2002): Assume $(T_i, \boldsymbol{C}_i)$ are consistent estimators for $(\boldsymbol{\mu}, a_i\boldsymbol{\Sigma})$ where $a_i > 0$ for $i = 1, 2$. Let $D_{i,1}$ and $D_{i,2}$ be the corresponding distances and let $R$ be the set of cases with distances $D_i(T_1, \boldsymbol{C}_1) \leq \text{MED}(D_i(T_1, \boldsymbol{C}_1))$. Let $r_n$ be the correlation between $D_{i,1}$ and $D_{i,2}$ for the cases in $R$. Then $r_n \to 1$ in probability as $n \to \infty$.

f) Olive (2004a): $(\overline{\boldsymbol{x}}_{0,50}, \boldsymbol{S}_{0,50})$ is a high breakdown estimator. If the data distribution is EC but not spherical about $\boldsymbol{\mu}$, then for $m \geq 0$, $\boldsymbol{S}_{m,50}$ 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. Arcones (1995) and Kim (2000) showed that $\overline{\boldsymbol{x}}_{0,50}$ is a HB $\sqrt{n}$ consistent estimator of $\boldsymbol{\mu}$.

The following remarks help explain why the FCH estimator is robust. Using $k = 5$ concentration steps often works well. The scaling makes $\boldsymbol{C}_{FCH}$

a better estimate of $\boldsymbol{\Sigma}$ if the data is multivariate normal MVN. See Equations (11.2) and (11.4). The attractor $(T_{k,0}, \boldsymbol{C}_{k,0})$ that uses the classical estimator (0% trimming) as a start is the DGK estimator and has good statistical properties. By Remark 10.4f, the start $(T_{0,50}, \boldsymbol{C}_{0,50})$ that uses 50% trimming is a high breakdown estimator. Since only cases $\boldsymbol{x}_i$ such that $\|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\| \leq \text{MED}(\|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\|)$ are used, the largest eigenvalue of $\boldsymbol{C}_{0,50}$ is bounded if fewer than half of the cases are outliers.

The geometric behavior of the start $(T_{0,50}, \boldsymbol{C}_{0,50})$ is simple. If the data $\boldsymbol{x}_i$ are MVN (or EC) then the highest density regions of the data are hyperellipsoids. The set of $\boldsymbol{x}$ closest to the coordinatewise median in Euclidean distance is a hypersphere. For EC data the highest density ellipsoid and hypersphere will have approximately the same center as the hypersphere, and the hypersphere will be drawn towards the longest axis of the hyperellipsoid. Hence too much data will be trimmed in that direction. For example, if the data are MVN with $\boldsymbol{\Sigma} = \text{diag}(1, 2, ..., p)$ then $\boldsymbol{C}_{0,50}$ will underestimate the largest variance and overestimate the smallest variance. Taking $k$ concentration steps can greatly reduce but not eliminate the bias of $\boldsymbol{C}_{k,50}$ if the data is EC, and the determinant $|\boldsymbol{C}_{k,50}| < |\boldsymbol{C}_{0,50}|$ unless the attractor is equal to the start by Remark 10.4c. The attractor $(T_{k,50}, \boldsymbol{C}_{k,50})$ is not affine equivariant but is resistant to gross outliers in that they will initially be given weight zero if they are further than the median Euclidean distance from the coordinatewise median. Gnanadesikan and Kettenring (1972, p. 94) suggest an estimator similar to the attractor $(T_{k,50}, \boldsymbol{C}_{k,50})$, also see Croux and Van Aelst (2002).

Recall that the sample median $\text{MED}(Y_i) = Y((n+1)/2)$ is the middle order statistic if $n$ is odd. Thus if $n = m + d$ where $m$ is the number of clean cases and $d = m - 1$ is the number of outliers so $\gamma \approx 0.5$, then the sample median can be driven to the max or min of the clean cases. The $j$th element of $\text{MED}(\boldsymbol{W})$ is the sample median of the $j$th predictor. Hence with $m-1$ outliers, $\text{MED}(\boldsymbol{W})$ can be driven to the "coordinatewise covering box" of the $m$ clean cases. The boundaries of this box are at the min and max of the clean cases from each predictor, and the lengths of the box edges equal the ranges $R_i$ of the clean cases for the $i$th variable. If $d \approx m/2$ so that $\gamma \approx 1/3$, then the $\text{MED}(\boldsymbol{W})$ can be moved to the boundary of the much smaller "coordinatewise IQR box" corresponding the 25th and 75th percentiles of the clean date. Then the edge lengths are approximately equal to the interquartile ranges of the clean cases.

Note that $D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p) = \|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\|$ is the Euclidean distance of $\boldsymbol{x}_i$ from MED($\boldsymbol{W}$). Let $\mathcal{C}$ denote the set of $m$ clean cases. If $d \leq m-1$, then the minimum distance of the outliers is larger than the maximum distance of the clean cases if the distances for the outliers satisfy $D_i > B$ where

$$B^2 = \max_{i \in \mathcal{C}} \|\boldsymbol{x}_i - \text{MED}(\boldsymbol{X})\|^2 \leq \sum_{i=1}^{p} R_i^2 \leq p(\max R_i^2).$$

**Example 10.4.** Tremearne (1911) recorded *height* = x[,1] and *height while kneeling* = x[,2] of 112 people. Figure 10.1a 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. The start $(\overline{\boldsymbol{x}}_{0,50}, \boldsymbol{S}_{0,50})$ is the classical estimator applied to the "half set" of cases closest to MED($\boldsymbol{W}$) in Euclidean distance. The circle (hypersphere for general $p$) 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 10.1b shows the DD plot of the classical distances vs 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.

As the dimension $p$ gets larger, outliers that can not be detected by marginal methods (case 44 in Example 10.4) become harder to detect. When $p = 3$ imagine that the clean data is a baseball bat 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 $p = 2$.

The MB estimator has outlier resistance similar to (MED($\boldsymbol{W}$), $\boldsymbol{I}_p$) for distant outliers but, as shown in Example 10.4, can be much more effective for detecting certain types of outliers that can not be found by marginal methods. For EC data, the MB estimator is best if the data is spherical about $\boldsymbol{\mu}$ or if the data is highly correlated with the major axis of the highest density region $\{\boldsymbol{x}_i : D_i^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leq d^2\}$.

Next, we will compare several concentration algorithms with theory and simulation. Let the CMCD algorithm use $k > 1$ concentration steps where

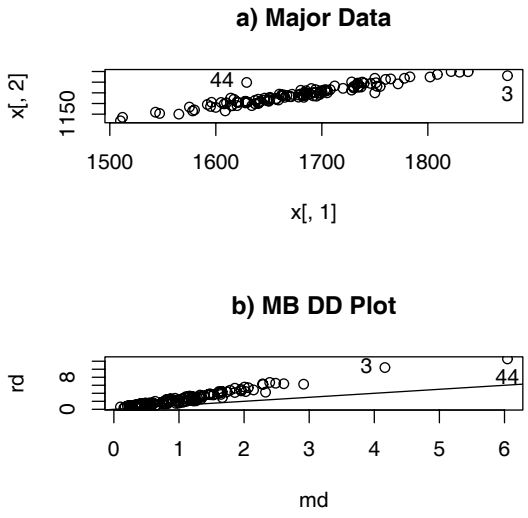


Figure 10.1: Plots for Major Data

the final estimator is the attractor that has the smallest determinant (the MCD criterion). We recommend $k = 10$ for the DGK estimator and $k = 5$ for the CMVE, FCH and MBA estimators.

To investigate the consistency and rate of robust estimators of multivariate location and dispersion, the following extension of Definitions 8.6 and 8.7 will be used. Let $g(n) \geq 1$ be an increasing function of the sample size $n$: $g(n) \uparrow \infty$, eg $g(n) = \sqrt{n}$. See White (1984, p. 15). Notice that if a $p \times 1$ random vector $T - \boldsymbol{\mu}$ converges to a nondegenerate multivariate normal distribution with convergence rate $\sqrt{n}$, then $T$ has (tightness) rate $\sqrt{n}$.

**Definition 10.14.** Let $\boldsymbol{A} = [a_{i,j}]$ be an $r \times c$ random matrix.
a) $\boldsymbol{A} = O_P(X_n)$ if $a_{i,j} = O_P(X_n)$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
b) $\boldsymbol{A} = o_p(X_n)$ if $a_{i,j} = o_p(X_n)$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
c) $\boldsymbol{A} \asymp_P (1/(g(n))$ if $a_{i,j} \asymp_P (1/(g(n))$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
d) Let $\boldsymbol{A}_1 = T - \boldsymbol{\mu}$ and $\boldsymbol{A}_2 = \boldsymbol{C} - c\boldsymbol{\Sigma}$ for some constant $c > 0$. If $\boldsymbol{A}_1 \asymp_P (1/(g(n))$ and $\boldsymbol{A}_2 \asymp_P (1/(g(n))$, then $(T, \boldsymbol{C})$ has (tightness) rate $g(n)$.

In MLR, if the start is a consistent estimator for $\boldsymbol{\beta}$, then so is the attractor. Hence all attractors are estimating the *same* parameter. The following proposition shows that MLD concentration estimators with $k \geq 1$ are esti-

mating $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$. Hence Remark 10.4 b) and d) can be combined with $d = a_{MCD}$ to provide simple proofs for MLD concentration algorithms.

**Proposition 10.12.** Assume that (E1) holds and that $(T, \boldsymbol{C})$ is a consistent estimator of for $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with rate $n^{\delta}$ where the constants $a > 0$ and $\delta > 0$. Then the classical estimator $(\overline{\boldsymbol{x}}_{m,j}, \boldsymbol{S}_{m,j})$ computed after trimming the $c_n \approx n/2$ of cases with the largest distances $D_i(T, \boldsymbol{C})$ is a consistent estimator for $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ with the same rate $n^{\delta}$. Hence $\text{MED}(D_i^2(\overline{\boldsymbol{x}}_{m,j}, \boldsymbol{S}_{m,j}))$ is a consistent estimator of $U_{0.5}/a_{MCD}$.

**Proof.** The result follows by Remark 10.4b if $a_{50} = a_{MCD}$. But by Remark 10.4e the overlap of cases used to compute $(\overline{\boldsymbol{x}}_{m,j}, \boldsymbol{S}_{m,j})$ and $(T_{MCD}, \boldsymbol{C}_{MCD})$ goes to 100% as $n \to \infty$. Hence the two sample covariance matrices $\boldsymbol{S}_{m,j}$ and $\boldsymbol{C}_{MCD}$ both estimate the same quantity $a_{MCD}\boldsymbol{\Sigma}$. QED

The following proposition proves that the elemental concentration and "h–set" basic resampling algorithms produce inconsistent zero breakdown estimators.

**Proposition 10.13.** Suppose that each start uses $h \geq p + 1$ randomly selected cases and that the number of starts $K_n \equiv K$ does not depend on $n$ (eg, $K = 500$). Then
i) the ("h-set") basic resampling estimator is inconsistent.
ii) The k–step CMCD concentration algorithm is inconsistent.
iii) For the basic resampling algorithm, the breakdown value is bounded above by $K/n$.
iv) For CMCD the breakdown value is bounded above by $K(h - p)/n$.

**Proof.** To prove i) and ii), notice that each start is inconsistent. Hence each attractor is inconsistent by Lopuhaä (1999) for CMCD. Choosing from $K$ inconsistent estimators still results in an inconsistent estimator. iii) Replace one case in each start by a case with a value tending to $\infty$. iv). If $h \geq p+1$, replace $h-p$ cases so that the start is singular and the covariance matrix can not be computed. QED

We certainly prefer to use consistent estimators whenever possible. When the start subset size $h_n \equiv h$ and the number of starts $K_n \equiv K$ are both fixed, the estimator is inconsistent. The situation changes dramatically if the start subset size $h_n = g(n) \to \infty$ as $n \to \infty$. The conditions in Proposition 10.14i hold, for example, if the classical estimator is applied to $h_n$ cases randomly

drawn from a distribution with a covariance matrix $\mathrm{Cov}(\boldsymbol{X}) = c_X\boldsymbol{\Sigma}$. Then each of the $K$ starts estimates $(\boldsymbol{\mu}, c_X\boldsymbol{\Sigma})$ with rate $[h_n]^{1/2}$.

**Proposition 10.14.** Suppose $K_n \equiv K$ starts are used and that all starts have subset size $h_n = g(n) \uparrow \infty$ as $n \to \infty$. Assume that the estimator applied to the subset has rate $n^\delta$.
i) If each of the $K$ estimators $(T_i, \boldsymbol{C}_i)$ is a $[g(n)]^\delta$ consistent estimator for $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ (ie, $a_i \equiv a$ for $i = 1, ..., K$), then the MLD $h_n$-set basic resampling algorithm estimator has rate $[g(n)]^\delta$.
ii) The CMCD estimator has rate $[g(n)]^\delta$ if assumption (E1) holds.
iii) The DGK estimator has rate $n^{1/2}$ if assumption (E1) holds.
iv) The MBA and FCH estimators have rate $n^{1/2}$ if (E1) holds and the distribution is spherical about $\boldsymbol{\mu}$.

**Proof.** i) The result follows by Pratt (1959). ii) By Lopuhaä (1999), all $K$ attractors have $[g(n)]^\delta$ rate, and the result follows by Proposition 10.12 and Pratt (1959). iii) The DGK estimator uses $K = 1$ and $h_n = n$, and the $k$ concentration steps are performed after using the classical estimator as a start. Hence the result follows by Lopuhaä (1999). iv) Each of the $K = 2$ starts is $\sqrt{n}$ consistent (if $M = 50$ then the $(\mathrm{MED}(\boldsymbol{W}), \boldsymbol{I}_p) = (T_{-1}, \boldsymbol{C}_{-1})$ can be regarded as the start). Hence the result follows by Proposition 10.12 and Pratt (1959). QED

Suppose that the concentration algorithm covers $c_n$ cases. Then Remark 10.3 suggested that concentration algorithms using $K$ starts each consisting of $h$ cases can handle roughly a percentage $\gamma_o$ of huge outliers where

$$\gamma_o \approx \min(\frac{n - c_n}{n}, 1 - [1 - (0.2)^{1/K}]^{1/h})100\% \tag{10.22}$$

if $n$ is large. Empirically, this value seems to give a rough approximation for many simulated data sets.

However, if the data set is multivariate and the bulk of the data falls in one compact ellipsoid while the outliers fall in another hugely distant compact ellipsoid, 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.

**Proposition 10.15.** Consider the CMCD and MCD estimators that both cover $c_n$ cases. For multivariate data, if at least one of the starts is nonsingular, then the CMCD estimator $\boldsymbol{C}_A$ 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 $0 \leq \det(\boldsymbol{C}_{MCD}) \leq \det(\boldsymbol{C}_A)$. QED

Next we will show that it is simple to modify existing elemental concentration algorithms such that the resulting CMCD estimators have good statistical properties. These CMCD estimators satisfy i) $0 < det(\boldsymbol{C}_{CMCD}) < \infty$ even if nearly half of the cases are outliers, and if (E1) holds then ii) $CMCD - MCD = O_P(n^{-1/2})$, and iii) the CMCD estimators are asymptotically equivalent to the DGK estimator if (E1) holds but the data distribution is not spherical about $\boldsymbol{\mu}$.

We will be interested in the attractor that minimizes the MCD criterion $det(\boldsymbol{S}_{k,M})$ and in the attractor that minimizes the MVE criterion

$$[MED(D_i)]^p \sqrt{det(\boldsymbol{S}_{k,M})}, \tag{10.23}$$

(see Rousseeuw and Leroy 1987, p. 259) which is proportional to the volume of the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - \overline{\boldsymbol{x}}_{k,M})^T \boldsymbol{S}_{k,M}^{-1} (\boldsymbol{z} - \overline{\boldsymbol{x}}_{k,M}) \leq d^2\} \tag{10.24}$$

where $d^2 = \text{MED}(D_i^2(\overline{\boldsymbol{x}}_{k,M}, \boldsymbol{S}_{k,M}))$. The following two theorems show how to produce $\sqrt{n}$ consistent robust estimators from starts that use $O(n)$ cases. The following theorem shows that the MBA and FCH estimators have good statistical properties.

**Theorem 10.16.** Suppose (E1) holds.

a) If $(T_A, \boldsymbol{C}_A)$ is the attractor that minimizes the MVE criterion (10.23), then $(T_A, \boldsymbol{C}_A)$ is a HB $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$.

b) If $(T_A, \boldsymbol{C}_A)$ is the attractor that minimizes $det(\boldsymbol{S}_{k,M})$, then $(T_A, \boldsymbol{C}_A)$ is a HB $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$. The MBA and FCH estimators are HB $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ where $c = 1$ for MVN data.

**Proof.** a) The estimator is HB since $(\overline{\boldsymbol{x}}_{0,50}, \boldsymbol{S}_{0,50})$ is a high breakdown estimator and hence has a bounded volume if up to nearly 50% of the cases are outliers. If the distribution is spherical about $\boldsymbol{\mu}$ then the result follows by Proposition 10.14iv. Otherwise, the hyperellipsoid corresponding to the highest density region has at least one major axis and at least one minor axis. The estimators with $M > 0$ trim too much data in the direction of the major axis and hence the resulting attractor is not estimating the highest density region. But the DGK estimator ($M = 0$) is estimating the highest density region. Thus the probability that the DGK estimator is the attractor that minimizes the volume goes to one as $n \to \infty$, and $(T_A, \boldsymbol{C}_A)$ is asymptotically equivalent to the DGK estimator $(T_{k,0}, \boldsymbol{C}_{k,0})$. QED

b) Under (E1) the FCH and MBA estimators are asymptotically equivalent since $\|T_{k,0} - \text{MED}(\boldsymbol{W})\| \to 0$ in probability. The estimator is HB since $0 < det(\boldsymbol{C}_{MCD}) \leq det(\boldsymbol{C}_A) \leq det(\boldsymbol{S}_{0,50}) < \infty$ if up to nearly 50% of the cases are outliers. If the distribution is spherical about $\boldsymbol{\mu}$ then the result follows by Proposition 10.14iv. Otherwise, the estimators with $M > 0$ trim to much data in the direction of the major axis and hence the resulting attractor is not

estimating the highest density region. Hence $\boldsymbol{S}_{k,M}$ is not estimating $a_{MCD}\boldsymbol{\Sigma}$. But the DGK estimator $\boldsymbol{S}_{k,0}$ is a $\sqrt{n}$ consistent estimator of $a_{MCD}\boldsymbol{\Sigma}$ and $\|\boldsymbol{C}_{MCD} - \boldsymbol{S}_{k,0}\| = O_P(n^{-1/2})$. 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 $(T_{k,0}, \boldsymbol{C}_{k,0})$. The scaling (10.19) makes $c = 1$ for MVN data. QED

The proof for CMVE is nearly identical: the CMVE volume is bounded by that of MVE (the minimum volume ellipsoid estimator) and MB, and the DGK estimator can be used to estimate the highest density minimum volume region while MB volume is too large for nonspherical EC distributions.

The following theorem shows that fixing the inconsistent zero breakdown elemental CMCD algorithm is simple. Just add the two FCH starts.

**Theorem 10.17.** Suppose that (E1) holds and that the CMCD algorithm uses $K_n \equiv K$ randomly selected elemental starts (eg, K = 500), the start $(T_{0,0}, \boldsymbol{C}_{0,0})$ and the start $(T_{0,50}, \boldsymbol{C}_{0,50})$. The elemental attractor $(\overline{\boldsymbol{x}}_{k,j}, \boldsymbol{S}_{k,j})$ or the DGK estimator $(T_{k,0}, \boldsymbol{C}_{k,0}) \equiv (\overline{\boldsymbol{x}}_{k,0}, \boldsymbol{S}_{k,0})$ is not used if

$$\|\overline{\boldsymbol{x}}_{k,j} - \mathrm{MED}(\boldsymbol{W})\| > \mathrm{MED}(D_i(\mathrm{MED}(\boldsymbol{W}), \boldsymbol{I}_p)). \tag{10.25}$$

Then this CMCD estimator is a HB $\sqrt{n}$ consistent estimator. If the EC distribution is not spherical about $\boldsymbol{\mu}$, then the CMCD estimator is asymptotically equivalent to the DGK estimator.

**Proof.** The estimator is HB since $0 < det(\boldsymbol{C}_{MCD}) \leq det(\boldsymbol{C}_{CMCD}) \leq det(\boldsymbol{S}_{0,50}) < \infty$ if up to nearly 50% of the cases are outliers. Notice that the DGK estimator $(T_{k,0}, \boldsymbol{C}_{k,0})$ is the attractor for $(T_{0,0}, \boldsymbol{C}_{0,0})$. Under (E1), the probability that the attractor from a randomly drawn elemental set gets arbitrarily close to the MCD estimator goes to zero as $n \to \infty$. But $DGK - MCD = O_P(n^{-1/2})$. Since the number of randomly drawn elemental sets $K$ does not depend on $n$, the probability that the DGK estimator has a smaller criterion value than that of the best elemental attractor also goes to one. Hence if the distribution is spherical about $\boldsymbol{\mu}$ then (with probability going to one) one of the FCH attractors will minimize the criterion value and the result follows. If (E1) holds and the distribution is not spherical about $\mu$, then the probability that the DGK attractor minimizes the determinant goes to one as $n \to \infty$, and $(T_{CMCD}, \boldsymbol{C}_{CMCD})$ is asymptotically equivalent to the DGK estimator $(T_{k,0}, \boldsymbol{C}_{k,0})$. Using the location criterion to eliminate attractors does not affect the results since under (E1), the probability that

$\|T_{k,0} - \text{MED}(\boldsymbol{W})\| \leq \text{MED}(D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p))$ goes to one. QED

**Definition 10.14.** Compute $D_i^2(T_F, \boldsymbol{C}_F)$ where $F$ is the MBA, FCH or CMVE estimator. i) Then compute the classical estimator from the cases with $D_i^2 \leq \chi^2_{p,0.975}$ and ii) scale for normality using the right hand side of (10.19). Repeat steps i) and ii). The resulting estimator is the *RMBA, RFCH or RCMVE estimator.*

**Theorem 10.18.** The RMBA, RFCH and RCMVE estimators are $\sqrt{n}$ consistent HB MLD estimators.

**Proof.** Since the MBA, FCH and CMVE estimators are $\sqrt{n}$ consistent and HB, so are the RMBA, RFCH and RCMVE estimators by Lopuhaä (1999). The reweighting step is commonly used and is known to not change the breakdown value, although the maximum amount of bias does change.

To compare $(T_{MBA}, \boldsymbol{C}_{MBA})$, $(T_{RMBA}, \boldsymbol{C}_{RMBA})$ and $(T_{FMCD}, \boldsymbol{C}_{FMCD})$, we used simulated data with $n = 100$ cases and computed the FMCD estimator with the *R/Splus* function `cov.mcd`. Initially the data sets had no outliers, and all 100 cases were MVN with zero mean vector and $\boldsymbol{\Sigma} = \text{diag}(1,2, ..., \text{p})$. We generated 500 runs of this data with $p = 4$. The averaged diagonal elements of $\boldsymbol{C}_{MBA}$ were 1.196, 2.223, 3.137 and 4.277. (In the simulations, the scale factor in Equation (10.19) appeared to be slightly too large for small $n$ but slowly converged to the correct factor as $n$ increased.) The averaged diagonal elements of $\boldsymbol{C}_{RMBA}$ were 1.002, 2.001, 2.951 and 4.005. The averaged diagonal elements of $\boldsymbol{C}_{FMCD}$ were 0.841, 1.655, 2.453, and 3.387. The approximation $1.2\boldsymbol{C}_{FMCD} \approx \boldsymbol{\Sigma}$ was good. For all three matrices, all off diagonal elements had average values less than 0.047 in magnitude.

Next data sets with 40% outliers were generated. The last 60 cases were MVN with zero mean vector and $\boldsymbol{\Sigma} = \text{diag}(1,2, ..., \text{p})$. The first 40 cases were MVN with the same $\boldsymbol{\Sigma}$, but the $p \times 1$ mean vector $\boldsymbol{\mu} = (10, 10\sqrt{2}, ..., 10\sqrt{p})^T$. We generated 500 runs of this data using $p = 4$. Shown below are the averages of $\boldsymbol{C}_{MBA}$, $\boldsymbol{C}_{RMBA}$ and $\boldsymbol{C}_{FMCD}$. Notice that $\boldsymbol{C}_{FMCD}$ performed extremely well while the $\boldsymbol{C}_{MBA}$ entries were over inflated by a factor of about 2 since the outliers inflate the scale factor $\text{MED}(D_i^2(T_A, \boldsymbol{C}_A))/\chi^2_{p,0.5}$.

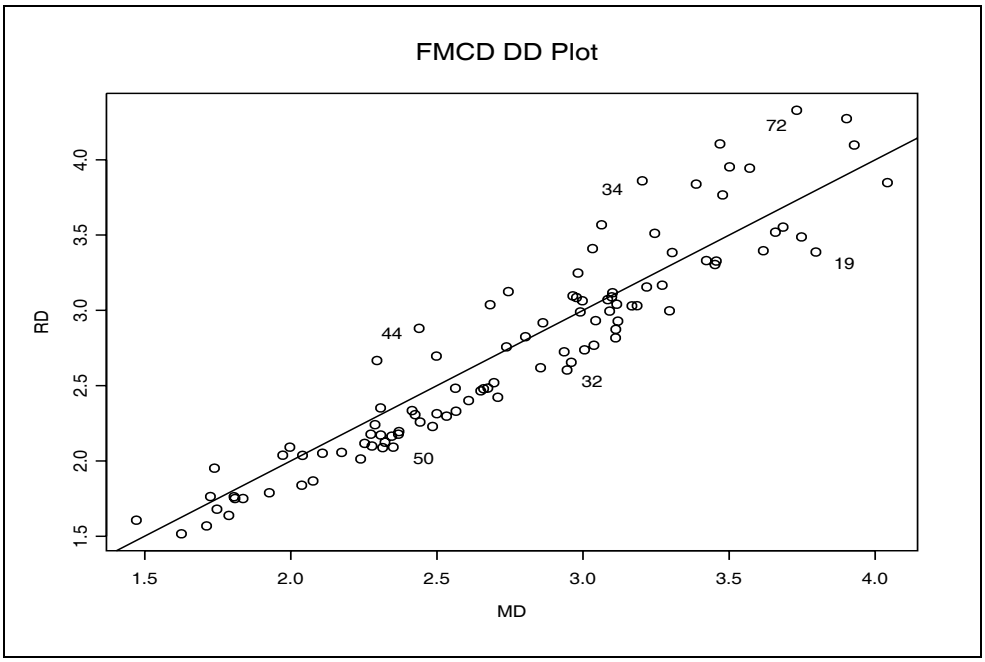


Figure 10.2: The FMCD Estimator Failed

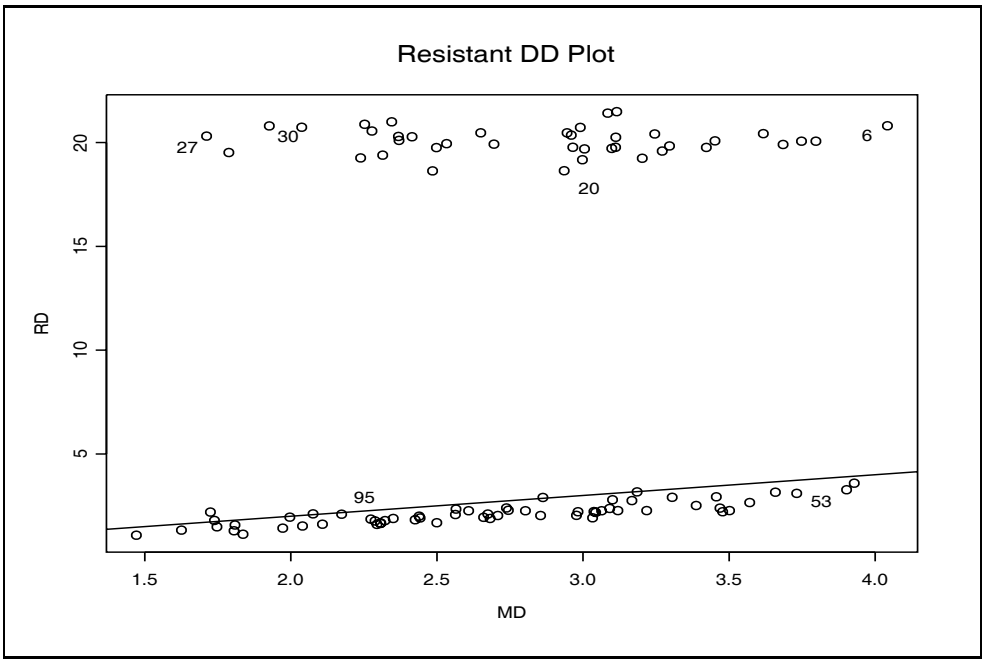


Figure 10.3: The Outliers are Large in the MBA DD Plot

MBA

$$\begin{bmatrix} 2.107 & -0.001 & 0.014 & -0.082 \\ -0.011 & 4.151 & -0.053 & -0.093 \\ 0.014 & -0.053 & 6.085 & -0.045 \\ -0.082 & -0.093 & -0.045 & 8.039 \end{bmatrix}$$

RMBA

$$\begin{bmatrix} 1.879 & 0.004 & -0.010 & -0.061 \\ 0.004 & 3.790 & 0.015 & 0.014 \\ -0.010 & 0.015 & 5.649 & 0.092 \\ -0.061 & 0.014 & 0.092 & 7.480 \end{bmatrix}$$

FMCD

$$\begin{bmatrix} 0.979 & 0.005 & -0.009 & -0.032 \\ 0.005 & 1.971 & 0.012 & 0.004 \\ -0.009 & 0.012 & 2.953 & 0.046 \\ -0.032 & 0.004 & 0.046 & 3.893 \end{bmatrix}$$

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 outliers in the MBA DD plot will have large $RD_i$ since $\boldsymbol{C}_{MBA} \approx 2\boldsymbol{C}_{FMCD} \approx 2\boldsymbol{\Sigma}$.

When $p$ is increased to 8, the `cov.mcd` estimator was usually not useful for detecting the outliers for this type of contamination. Figure 10.2 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 10.3.

**Remark 10.5.** Assume assumption (E1) holds, and consider modifying the FMCD algorithm by adding the 2 MBA starts. The FMCD estimator uses 500 elemental starts and partitioning and also iterates 5 starts to convergence. Suppose the data set has $n_D$ cases. Then the maximum number of concentration steps until convergence is bounded by $k_D$, say. Assume that for $n > n_D$, no more than $k_D$ concentration steps are used. (This assumption is not unreasonable. Asymptotic theory is meant to simplify matters, not to make things more complex. Also the algorithm is supposed to be fast. Letting the maximum number of concentration steps increase to $\infty$ would result in an impractical algorithm.) Then the elemental attractors are inconsistent and for EC data that is not spherical about $\boldsymbol{\mu}$, the best attractor will be asymptotically equivalent with the DGK estimator. The modified FMCD

"weight for efficiency step" does not change the $\sqrt{n}$ rate by Lopuhaä (1999). The algorithm can be further improved by not using attractors satisfying Equation (10.25).

A simple simulation for outlier resistance is to generate outliers and count the percentage of times the minimum distance of the outliers is larger than the maximum distance of the clean cases. Then the outliers can be separated from the clean cases with a horizontal line in the DD plot. The simulation used 100 runs and $n = 200$. If $\gamma = 0.2$ then the first 40 cases were outliers. The clean cases were MVN: $\boldsymbol{x} \sim N_p(\mathbf{0}, diag(1, 2, ..., p))$. Outlier types were 1) a point mass $(0, ..., 0, pm)^T$ at the major axis, 2) a point mass $(pm, 0, ..., 0)^T$ at the minor axis and 3) $\boldsymbol{x} \sim N_p(pm\mathbf{1}, diag(1, 2, ..., p))$ where $\mathbf{1} = (1, ..., 1)^T$.

Maronna and Zamar (2002) suggest that a point mass orthogonal to the major axis may be least favorable for OGK, but for FAST-MCD and MBA a point mass at the major axis will cause a lot of difficulty because an ellipsoid with very small volume can cover half of the data by putting the outliers at one end of the ellipsoid and the clean data in the other end. This half set will produce a classical estimator with very small determinant by (10.23). Rocke and Woodruff (1996) suggest that outliers with a mean shift are hard to detect. A point mass is used although for large $\gamma$ and moderate $p$ the point mass causes numerical difficulties in that the $R$ software will declare that the sample covariance matrix is singular.

Notice that the clean data can be transformed to a $N_p(\mathbf{0}, \boldsymbol{I}_p)$ distribution by multiplying $\boldsymbol{x}_i$ by $\text{diag}(1, 1/\sqrt{2}, ..., 1/\sqrt{p})$. The counts for affine equivariant estimators such as DGK and FAST-MCD will not be changed. Notice that the point mass at the minor axis $(pm, 0, ..., 0)^T$ is not changed by the transformation, but the point mass at the major axis becomes $(0, ..., 0, pm/\sqrt{p})^T$, which is much harder to detect.

The results of the simulation are shown in Table 10.1. The counts for the classical estimator were always 0 and thus omitted. As expected, the MCD criterion has trouble with a tight cluster of outliers. For $p = 20$, $\gamma = .2$ and a point mass at the major axis, FAST-MCD needed PM = 4000 and MBA needed PM = 10000 before having a small chance of giving the outliers large distances. Combining information from location and dispersion was effective. The point mass outliers make the DGK determinant small (though larger than the MCD determinant by definition), but pull the DGK location estimator away from MED($\boldsymbol{W}$). Note that FCH performance dominated MBA and was sometimes better than OGK and sometimes worse. CMVE

Table 10.1: Percentage of Times Outliers Were Detected

| p | $\gamma$ | type | PM | MBA | FCH | DGK | OGK | FMCD | CMVE | MB |
|---|---|---|---|---|---|---|---|---|---|---|
| 5 | .2 | 1 | 15 | 0 | 100 | 0 | 0 | 0 | 100 | 100 |
| 10 | .2 | 1 | 20 | 0 | 4 | 0 | 0 | 0 | 16 | 96 |
| 20 | .2 | 1 | 30 | 0 | 0 | 0 | 0 | 0 | 1 | 61 |
| 20 | .2 | 1 | 50 | 0 | 100 | 0 | 0 | 0 | 100 | 100 |
| 20 | .2 | 1 | 100 | 0 | 100 | 0 | 22 | 0 | 100 | 100 |
| 20 | .2 | 1 | 4000 | 0 | 100 | 0 | 100 | 31 | 100 | 100 |
| 20 | .2 | 1 | 10000 | 24 | 100 | 0 | 100 | 100 | 100 | 100 |
| 5 | .2 | 2 | 15 | 97 | 100 | 0 | 71 | 100 | 100 | 100 |
| 10 | .2 | 2 | 20 | 0 | 58 | 0 | 71 | 0 | 97 | 100 |
| 20 | .2 | 2 | 30 | 0 | 0 | 0 | 99 | 0 | 76 | 100 |
| 20 | .2 | 2 | 50 | 0 | 100 | 0 | 100 | 0 | 100 | 100 |
| 20 | .2 | 2 | 100 | 0 | 100 | 0 | 100 | 0 | 100 | 100 |
| 20 | .2 | 2 | 4000 | 96 | 100 | 0 | 100 | 100 | 100 | 100 |
| 5 | .2 | 3 | 5 | 88 | 88 | 87 | 5 | 97 | 92 | 91 |
| 10 | .2 | 3 | 5 | 92 | 92 | 84 | 2 | 100 | 92 | 94 |
| 20 | .2 | 3 | 5 | 85 | 85 | 1 | 0 | 99 | 66 | 85 |
| 40 | .4 | 3 | 20 | 38 | 38 | 0 | 0 | 0 | 40 | 100 |
| 40 | .4 | 3 | 30 | 77 | 97 | 0 | 59 | 0 | 91 | 100 |
| 40 | .4 | 3 | 40 | 91 | 100 | 0 | 100 | 0 | 100 | 100 |

was nearly always better than OGK. For a mean shift and small $p$ and $\gamma$ the elemental FAST-MCD estimator was somewhat better than CMVE, MB, MBA and FCH. If $\gamma$ is large enough then CMVE, MBA, FCH and MB dominate FAST-MCD. MB was never worse than OGK, but OGK did seem to behave like a HB estimator in that it could detect distant outliers.

The simulation suggests that marginal methods for detecting outliers should not be abandoned. We suggest making a DD plot with the $\sqrt{n}$ consistent HB FCH estimator as an EC diagnostic. Make the MB DD plot to check for outliers. Other methods that do not have proven theory can also be used as outlier diagnostics. For $p \leq 10$ make a scatterplot matrix of the variables. The plots should be ellipsoidal if the EC assumption is reasonable. Dot plots of individual predictors with superimposed histograms are

also useful. For large $n$ the histograms should be approximately symmetric if the EC assumption is reasonable.

**Software**

The `robustbase` library was downloaded from (www.r-project.org/#doc). § 14.2 explains how to use the source command to get the `rpack` functions in $R$ and how to download a library from $R$. Type the commands `library(MASS)` and `library(robustbase)` to compute the FAST-MCD and OGK estimators with the `cov.mcd` and `covOGK` functions.

The `rpack` function
*mldsim(n=200,p=5,gam=.2,runs=100,outliers=1,pm=15)*
can be used to simulate the first line in Table 10.1. Change outliers to 0 to examine the average of $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$. The function `mldsim5` is similar but does not need the `library` command since it compares the FCH, RFCF, CMVE, RCMVE and MB estimators. The command
*sctplt(n=200,p=10,gam=.2,outliers=3, pm=5)*
will make 1 data set corresponding to 5th line from the bottom of Table 10.1. 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 much faster than OGK which is much faster than FAST-MCD. The function *corrsim* can be used to simulate the correlations of robust distances with classical distances. The RCMVE, RMBA and RFCH are reweighted versions of CMVE, MBA and FCH that may perform better for small $n$. 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.

Functions *covdgk*, *covmba* and *rmba* compute the scaled DGK, MBA and RMBA estimators while *covfch* and *cmve* are used to compute FCH, RFCH, CMVE and RCMVE.

## 10.8 Complements

The theory for concentration algorithms is due to Hawkins and Olive (2002) and Olive and Hawkins (2007b,2008). 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. An important application of the robust algorithm estimators and of case diagnostics is to detect outliers. Sometimes it can be assumed that the analysis for influential cases and outliers was completely successful in classifying the cases into outliers and good or "clean" cases. Then classical procedures can be performed on the good cases. This assumption of perfect classification is often unreasonable, and it is useful to have robust procedures, such as the FCH estimator, that have rigorous asymptotic theory and are practical to compute. Since the FCH estimator is about an order of magnitude faster than alternative robust estimators, the FCH estimator may be useful for computationally intensive applications.

The RFCH estimator takes slightly longer to compute than the FCH estimator, and should have slightly less resistance to outliers.

In addition to concentration and randomly selecting elemental sets, three other algorithm techniques are important. He and Wang (1996) suggest computing the classical estimator and a consistent robust estimator. The final cross checking estimator is the classical estimator if both estimators are "close," otherwise the final estimator is the robust estimator. The second technique was proposed by Gnanadesikan and Kettenring (1972, p. 90). They suggest using the dispersion matrix $\boldsymbol{C} = [c_{i,j}]$ where $c_{i,j}$ is a robust estimator of the covariance of $X_i$ and $X_j$. Computing the classical estimator on a subset of the data results in an estimator of this form. The identity

$$c_{i,j} = \mathrm{Cov}(\mathrm{X_i}, \mathrm{X_j}) = [\mathrm{VAR}(\mathrm{X_i} + \mathrm{X_j}) - \mathrm{VAR}(\mathrm{X_i} - \mathrm{X_j})]/4$$

where $\mathrm{VAR(X)} = \sigma^2(\mathrm{X})$ suggests that a robust estimator of dispersion can be created by replacing the sample standard deviation $\hat{\sigma}$ by a robust estimator of scale. Maronna and Zamar (2002) modify this idea to create a fairly fast high breakdown consistent OGK estimator of multivariate location and dispersion. This estimator may be the leading competitor of the FCH estimator. Also see Alqallaf, Konis, Martin and Zamar (2002) and Mehrotra (1995). Woodruff and Rocke (1994) introduced the third technique, partitioning, which evaluates a start on a subset of the cases. Poor starts are

discarded, and $L$ of the best starts are evaluated on the entire data set. This idea is also used by Rocke and Woodruff (1996) and by Rousseeuw and Van Driessen (1999).

There certainly exist types of outlier configurations where the FMCD estimator outperforms the robust FCH estimator. The FCH estimators is vulnerable to outliers that lie inside the hypersphere based on the median Euclidean distance from the coordinatewise median. Although the FCH estimator should not be viewed as a replacement for the FMCD estimator, the FMCD estimator should be modified as in Theorem 10.17. Until this modification appears in the software, both estimators can be used for outlier detection by making a scatterplot matrix of the Mahalanobis distances from the FMCD, FCH and classical estimators.

The simplest version of the MBA estimator only has two starts. A simple modification would be to add additional starts as in Problem 10.18.

Johnson and Wichern (1988) and Mardia, Kent and Bibby (1979) are good references for multivariate statistical analysis based on the multivariate normal distribution. The elliptically contoured distributions generalize the multivariate normal distribution and are discussed (in increasing order of difficulty) in Johnson (1987), Fang, Kotz and Ng (1990), Fang and Anderson (1990), and Gupta and Varga (1993). Fang, Kotz and Ng (1990) sketch the history of elliptically contoured distributions while Gupta and Varga (1993) discuss matrix valued elliptically contoured distributions. Cambanis, Huang and Simons (1981), Chmielewski (1981) and Eaton (1986) are also important references. Also see Muirhead (1982, p. 30–42).

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). Bernholdt and Fisher (2004) show that the MCD estimator can be computed with $O(n^v)$ complexity where $v = 1 + p(p+3)/2$ if $\boldsymbol{x}$ is a $p \times 1$ vector.

Rocke and Woodruff (1996, p. 1050) claim that any affine equivariant location and shape estimation method gives an unbiased location estimator and a shape estimator that has an expectation that is a multiple of the true shape for elliptically contoured distributions. Hence there are many candidate robust estimators of multivariate location and dispersion. See Cook, Hawkins and Weisberg (1993) for an exact algorithm for the MVE. Other papers on robust algorithms include Hawkins (1993b, 1994), Hawkins

and Olive (1999a), Hawkins and Simonoff (1993), He and Wang (1996), Olive (2004a), Olive and Hawkins (2007b, 2008), Rousseeuw and Van Driessen (1999), Rousseeuw and van Zomeren (1990), Ruppert (1992), and Woodruff and Rocke (1993). Rousseeuw and Leroy (1987, § 7.1) also describes many methods.

The discussion by Rocke and Woodruff (2001) and by Hubert (2001) of Peña and Prieto (2001) stresses the fact that no one estimator can dominate all others for every outlier configuration. These papers and Wisnowski, Simpson, and Montgomery (2002) give outlier configurations that can cause problems for the FMCD estimator.

Papers on robust distances include Olive (2002) and García-Escudero and Gordaliza (2005).

## 10.9 Problems

**10.1*.** 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 $\rho(X_1, X_3)$.

**10.2*.** 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 $\sigma_{12} = \mathrm{Cov}(Y, X)$ 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 $\text{Var}(Y|X)$.

**10.3.** 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 $X$.

**10.4.** 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 10.2, show that $\boldsymbol{X}$ has an elliptically contoured distribution assuming that all relevant expectations exist.

**10.5.** In Proposition 10.5b, show that if the second moments exist, then $\boldsymbol{\Sigma}$ can be replaced by $\text{Cov}(\boldsymbol{X})$.

| crancap | hdlen | hdht | Data for 10.6 |
|---|---|---|---|
| 1485 | 175 | 132 | |
| 1450 | 191 | 117 | |
| 1460 | 186 | 122 | |
| 1425 | 191 | 125 | |
| 1430 | 178 | 120 | |
| 1290 | 180 | 117 | |
| 90 | 75 | 51 | |

**10.6*.** 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 $\overline{\boldsymbol{x}}$.

**10.7.** Using the notation in Proposition 10.6, show that if the second moments exist, then

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

**10.8.** Using the notation under Lemma 10.4, 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.

**10.9*.** 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.

**10.10.** Recall that $\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = E[(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{Y} - E(\boldsymbol{Y}))^T]$. Using the notation of Proposition 10.6, 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

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

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

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

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

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

**R/Splus Problems**

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

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

b) Enter the following commands and check whether observations 1–40 look like outliers.

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

**10.13.** Download the `rmaha` function that creates the robust Mahalanobis distances. Obtain outx2 as in Problem 10.12 b). *R* users need to enter the command *library(MASS)*. Enter the command *rmaha(outx2)* and check whether observations 1–40 look like outliers.

**10.14.** 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.

**10.15*.** a) Assuming that you have done the two source commands above Problem 10.12 (and in *R* the library(MASS) command), 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 each 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 the 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 (1905-6) 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.

**10.16*.** (Perform the *source("A:/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 median ball, 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. 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*.

**10.17*.** (Perform the *source("A:/rpack.txt")* command if you have not already done so.) The *ddmv* function illustrates concentration with the DD plot. 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.

**10.18.** (Perform the *source("A:/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 10.15.