

## Chapter 4

# MLD Estimators

This chapter is the most important chapter for outlier robust statistics and covers robust estimators of multivariate location and dispersion. The practical, highly outlier resistant,  $\sqrt{n}$  consistent FCH, RFCH, and RMVN estimators of  $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$  are developed along with proofs. The RFCH and RMVN estimators are reweighted versions of the FCH estimator. It is shown why competing “robust estimators” fail to work, are impractical, or are not yet backed by theory. The RMVN and RFCH sets are defined and will be used to create practical robust methods of principal component analysis, canonical correlation analysis, discriminant analysis, factor analysis, and multivariate linear regression in the following chapters.

**Warning:** This chapter contains many acronyms, abbreviations, and estimator names such as FCH, RFCH, and RMVN. See Section 1.2 and Table 1.1 for many of the acronyms and for the acronyms that start with the added letter A, C, F, or R: A stands for *algorithm*, C for *concentration*, F for estimators that use a *fixed* number of trial fits, and R for *reweighted*.

Let  $\boldsymbol{\mu}$  be a  $p \times 1$  location vector and  $\boldsymbol{\Sigma}$  a  $p \times p$  symmetric dispersion matrix. Because of symmetry, the first row of  $\boldsymbol{\Sigma}$  has  $p$  distinct unknown parameters, the second row has  $p-1$  distinct unknown parameters, the third row has  $p-2$  distinct unknown parameters, ..., and the  $p$ th row has one distinct unknown parameter for a total of  $1+2+\cdots+p = p(p+1)/2$  unknown parameters. Since  $\boldsymbol{\mu}$  has  $p$  unknown parameters, an estimator  $(T, \mathbf{C})$  of multivariate location and dispersion (MLD), needs to estimate  $p(p+3)/2$  unknown parameters when there are  $p$  random variables. If the  $p$  variables can be transformed into an uncorrelated set then there are only  $2p$  parameters, the means and variances, while if the dimension can be reduced from  $p$  to  $p-1$ , the number of parameters is reduced by  $p(p+3)/2 - (p-1)(p+2)/2 = p+1$ .

The sample covariance or sample correlation matrices estimate these parameters very efficiently since  $\boldsymbol{\Sigma} = (\sigma_{ij})$  where  $\sigma_{ij}$  is a population covariance or correlation. These quantities can be estimated with the sample covariance

or correlation taking two variables  $X_i$  and  $X_j$  at a time. Note that there are  $p(p+1)/2$  pairs that can be chosen from  $p$  random variables  $X_1, \dots, X_p$ .

**Rule of thumb 4.1.** For the classical estimators of multivariate location and dispersion,  $(\bar{\mathbf{x}}, \mathbf{S})$  or  $(\bar{\mathbf{z}} = \mathbf{0}, \mathbf{R})$ , we want  $n \geq 10p$ . We want  $n \geq 20p$  for the robust MLD estimators (FCH, RFCH, or RMVN) described later in this chapter.

## 4.1 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  $\mathbf{W}$ . Let  $\mathbf{B} = \mathbf{1}\mathbf{b}^T$  where  $\mathbf{1}$  is an  $n \times 1$  vector of ones and  $\mathbf{b}$  is a  $p \times 1$  constant vector. Hence the  $i$ th row of  $\mathbf{B}$  is  $\mathbf{b}_i^T \equiv \mathbf{b}^T$  for  $i = 1, \dots, n$ . For such a matrix  $\mathbf{B}$ , consider the affine transformation  $\mathbf{Z} = \mathbf{W}\mathbf{A}^T + \mathbf{B}$  where  $\mathbf{A}$  is any nonsingular  $p \times p$  matrix. An affine transformation changes  $\mathbf{x}_i$  to  $\mathbf{z}_i = \mathbf{A}\mathbf{x}_i + \mathbf{b}$  for  $i = 1, \dots, n$ , and affine equivariant multivariate location and dispersion estimators change in natural ways.

**Definition 4.1.** The multivariate location and dispersion estimator  $(T, \mathbf{C})$  is *affine equivariant* if

$$T(\mathbf{Z}) = T(\mathbf{W}\mathbf{A}^T + \mathbf{B}) = \mathbf{A}T(\mathbf{W}) + \mathbf{b}, \quad (4.1)$$

$$\text{and } \mathbf{C}(\mathbf{Z}) = \mathbf{C}(\mathbf{W}\mathbf{A}^T + \mathbf{B}) = \mathbf{A}\mathbf{C}(\mathbf{W})\mathbf{A}^T. \quad (4.2)$$

The following proposition shows that the Mahalanobis distances are invariant under affine transformations. See Rousseeuw and Leroy (1987, pp. 252–262) for similar results. Thus if  $(T, \mathbf{C})$  is affine equivariant, so is  $(T, D_{(c_n)}^2(T, \mathbf{C}) \mathbf{C})$  where  $D_{(j)}^2(T, \mathbf{C})$  is the  $j$ th order statistic of the  $D_i^2$ .

**Proposition 4.1.** If  $(T, \mathbf{C})$  is affine equivariant, then

$$D_i^2(\mathbf{W}) \equiv D_i^2(T(\mathbf{W}), \mathbf{C}(\mathbf{W})) = D_i^2(T(\mathbf{Z}), \mathbf{C}(\mathbf{Z})) \equiv D_i^2(\mathbf{Z}). \quad (4.3)$$

**Proof.** Since  $\mathbf{Z} = \mathbf{W}\mathbf{A}^T + \mathbf{B}$  has  $i$ th row  $\mathbf{z}_i^T = \mathbf{x}_i^T \mathbf{A}^T + \mathbf{b}^T$ ,

$$\begin{aligned} D_i^2(\mathbf{Z}) &= [\mathbf{z}_i - T(\mathbf{Z})]^T \mathbf{C}^{-1}(\mathbf{Z}) [\mathbf{z}_i - T(\mathbf{Z})] \\ &= [\mathbf{A}(\mathbf{x}_i - T(\mathbf{W}))]^T [\mathbf{A}\mathbf{C}(\mathbf{W})\mathbf{A}^T]^{-1} [\mathbf{A}(\mathbf{x}_i - T(\mathbf{W}))] \\ &= [\mathbf{x}_i - T(\mathbf{W})]^T \mathbf{C}^{-1}(\mathbf{W}) [\mathbf{x}_i - T(\mathbf{W})] = D_i^2(\mathbf{W}). \quad \square \end{aligned}$$

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

## 4.2 Breakdown

This section gives a standard definition of breakdown for estimators of multivariate location and dispersion. The following notation will be useful. Let  $\mathbf{W}$  denote the  $n \times p$  data matrix with  $i$ th row  $\mathbf{x}_i^T$  corresponding to the  $i$ th case. Let  $\mathbf{w}_1, \dots, \mathbf{w}_n$  be the contaminated data after  $d_n$  of the  $\mathbf{x}_i$  have been replaced by arbitrarily bad contaminated cases. Let  $\mathbf{W}_d^n$  denote the  $n \times p$  data matrix with  $i$ th row  $\mathbf{w}_i^T$ . Then the contamination fraction is  $\gamma_n = d_n/n$ . Let  $(T(\mathbf{W}), \mathbf{C}(\mathbf{W}))$  denote an estimator of multivariate location and dispersion where the  $p \times 1$  vector  $T(\mathbf{W})$  is an estimator of location and the  $p \times p$  symmetric positive semidefinite matrix  $\mathbf{C}(\mathbf{W})$  is an estimator of dispersion. Recall from Theorem 1.1 that if  $\mathbf{C}(\mathbf{W}_d^n) > 0$ , then  $\max_{\|\mathbf{a}\|=1} \mathbf{a}^T \mathbf{C}(\mathbf{W}_d^n) \mathbf{a} = \lambda_1$  and

$\min_{\|\mathbf{a}\|=1} \mathbf{a}^T \mathbf{C}(\mathbf{W}_d^n) \mathbf{a} = \lambda_p$ . A high breakdown dispersion estimator  $\mathbf{C}$  is positive definite if the amount of contamination is less than the breakdown value. Since  $\mathbf{a}^T \mathbf{C} \mathbf{a} = \sum_{i=1}^p \sum_{j=1}^p c_{ij} a_i a_j$ , the largest eigenvalue  $\lambda_1$  is bounded as  $\mathbf{W}_d^n$  varies iff  $\mathbf{C}(\mathbf{W}_d^n)$  is bounded as  $\mathbf{W}_d^n$  varies.

**Definition 4.2.** The *breakdown value* of the multivariate location estimator  $T$  at  $\mathbf{W}$  is

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

where the supremum is over all possible corrupted samples  $\mathbf{W}_d^n$  and  $1 \leq d_n \leq n$ . Let  $\lambda_1(\mathbf{C}(\mathbf{W})) \geq \dots \geq \lambda_p(\mathbf{C}(\mathbf{W})) \geq 0$  denote the eigenvalues of the dispersion estimator applied to data  $\mathbf{W}$ . The estimator  $\mathbf{C}$  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

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

**Definition 4.3.** Let  $\gamma_n$  be the breakdown value of  $(T, \mathbf{C})$ . *High breakdown (HB) statistics* have  $\gamma_n \rightarrow 0.5$  as  $n \rightarrow \infty$  if the (uncontaminated) clean data are in *general position*: no more than  $p$  points of the clean data lie on any  $(p-1)$ -dimensional hyperplane. Estimators are *zero breakdown* if  $\gamma_n \rightarrow 0$  and *positive breakdown* if  $\gamma_n \rightarrow \gamma > 0$  as  $n \rightarrow \infty$ .

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

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

**Proposition 4.2.** Fix  $n$ . If nonequivariant estimators (that may have a breakdown value of greater than  $1/2$ ) are excluded, then a multivariate location estimator has a breakdown value of  $d_T/n$  iff  $d_T = d_{T,n}$  is the smallest number of arbitrarily bad cases that can make the median Euclidean distance  $\text{MED}(\|\mathbf{w}_i - T(\mathbf{W}_d^n)\|)$  arbitrarily large.

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

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

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

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

**Lemma 4.3.** If the classical estimator  $(\bar{\mathbf{X}}_B, \mathbf{S}_B)$  is applied to  $c_n$  cases that are contained in some bounded region where  $p + 1 \leq c_n \leq n$ , then the maximum eigenvalue  $\lambda_1$  of  $\mathbf{S}_B$  is bounded.

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

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

Hence the maximum eigenvalue  $\lambda_1$  is bounded.  $\square$

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

$$\{\mathbf{z} : (\mathbf{z} - T)^T \mathbf{C}^{-1} (\mathbf{z} - T) \leq D_{(c_n)}^2\} \quad (4.4)$$

where  $D_{(c_n)}^2$  is the  $c_n$ th smallest squared Mahalanobis distance based on  $(T, \mathbf{C})$ . This hyperellipsoid contains the  $c_n$  cases with the smallest  $D_i^2$ . Suppose  $(T, \mathbf{C}) = (\bar{\mathbf{x}}_M, b \mathbf{S}_M)$  is the sample mean and scaled sample covariance matrix applied to some subset of the data where  $b > 0$ . The classical, RFCH, and RMVN estimators satisfy this assumption. For  $h > 0$ , the hyperellipsoid

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

has volume equal to

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

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

### 4.3 The Concentration Algorithm

Concentration algorithms are widely used since impractical brand name estimators, such as the MCD estimator given in Definition 4.4, take too long to compute. The concentration algorithm, defined in Definition 4.5, use  $K$  starts and attractors. A *start* is an initial estimator, and an *attractor* is an estimator obtained by refining the start. For example, let the start be the

classical estimator  $(\bar{\mathbf{x}}, \mathbf{S})$ . Then the attractor could be the classical estimator  $(T_1, \mathbf{C}_1)$  applied to the half set of cases with the smallest Mahalanobis distances. This concentration algorithm uses one concentration step, but the process could be iterated for  $k$  concentration steps, producing an estimator  $(T_k, \mathbf{C}_k)$ .

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

The remainder of this chapter will explain the concentration algorithm, explain why the MCD criterion is useful but can be improved, provide some theory for practical robust multivariate location and dispersion estimators, and show how the set of cases used to compute the recommended RMVN or RFCH estimator can be used to create robust multivariate analogs of methods such as principal component analysis and canonical correlation analysis. The RMVN and RFCH estimators are reweighted versions of the practical FCH estimator, given in Definition 4.8.

**Definition 4.4.** 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  $\mathbf{C}_{MCD}$  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  $(T_{MCD}(\mathbf{W}), \mathbf{C}_{MCD}(\mathbf{W}))$ .

Here

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

is the binomial coefficient.

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

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

in the location model where LTS stands for the least trimmed sum of squares estimator. See Chapter 14. The population analog of the MCD estimator is closely related to the hyperellipsoid of highest concentration that contains  $c_n/n \approx$  half of the mass. The MCD estimator is a  $\sqrt{n}$  consistent HB asymptotically normal estimator for  $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$  where  $a_{MCD}$  is some positive constant when the data  $\mathbf{x}_i$  are iid from a large class of distributions. See Cator and Lopuhaä (2010, 2012) who extended some results of Butler et al. (1993).

Computing robust covariance estimators can be very expensive. For example, to compute the exact  $MCD(c_n)$  estimator  $(T_{MCD}, \mathbf{C}_{MCD})$ , we need to consider the  $C(n, c_n)$  subsets of size  $c_n$ . Woodruff and Rocke (1994, p. 893)

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

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

**Definition 4.5.** Suppose that  $\mathbf{x}_1, \dots, \mathbf{x}_n$  are  $p \times 1$  vectors of observed data. For the multivariate location and dispersion model, an *elemental set*  $J$  is a set of  $p + 1$  cases. An elemental start is the sample mean and sample covariance matrix of the data corresponding to  $J$ . In a *concentration algorithm*, let  $(T_{-1,j}, \mathbf{C}_{-1,j})$  be the  $j$ th start (not necessarily elemental) and compute all  $n$  Mahalanobis distances  $D_i(T_{-1,j}, \mathbf{C}_{-1,j})$ . At the next iteration, the classical estimator  $(T_{0,j}, \mathbf{C}_{0,j}) = (\bar{\mathbf{x}}_{0,j}, \mathbf{S}_{0,j})$  is computed from the  $c_n \approx n/2$  cases corresponding to the smallest distances. This iteration can be continued for  $k$  *concentration steps* resulting in the sequence of estimators  $(T_{-1,j}, \mathbf{C}_{-1,j}), (T_{0,j}, \mathbf{C}_{0,j}), \dots, (T_{k,j}, \mathbf{C}_{k,j})$ . The result of the iteration  $(T_{k,j}, \mathbf{C}_{k,j})$  is called the  $j$ th *attractor*. If  $K_n$  starts are used, then  $j = 1, \dots, K_n$ . The *concentration attractor*,  $(T_A, \mathbf{C}_A)$ , is the attractor chosen by the algorithm. The attractor is used to obtain the final estimator. A common choice is the attractor that has the smallest determinant  $\det(\mathbf{C}_{k,j})$ . The *basic resampling algorithm* estimator is a special case where  $k = -1$  so that the attractor is the start:  $(\bar{\mathbf{x}}_{k,j}, \mathbf{S}_{k,j}) = (\bar{\mathbf{x}}_{-1,j}, \mathbf{S}_{-1,j})$ .

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. The following proposition is useful and shows that  $\det(\mathbf{S}_{0,j})$  tends to be greater than the determinant of the attractor  $\det(\mathbf{S}_{k,j})$ .

**Proposition 4.4: Rousseeuw and Van Driessen (1999, p. 214).** Suppose that the classical estimator  $(\bar{\mathbf{x}}_{t,j}, \mathbf{S}_{t,j})$  is computed from  $c_n$  cases and that the  $n$  Mahalanobis distances  $D_i \equiv D_i(\bar{\mathbf{x}}_{t,j}, \mathbf{S}_{t,j})$  are computed. If  $(\bar{\mathbf{x}}_{t+1,j}, \mathbf{S}_{t+1,j})$  is the classical estimator computed from the  $c_n$  cases with the smallest Mahalanobis distances  $D_i$ , then  $\det(\mathbf{S}_{t+1,j}) \leq \det(\mathbf{S}_{t,j})$  with equality iff  $(\bar{\mathbf{x}}_{t+1,j}, \mathbf{S}_{t+1,j}) = (\bar{\mathbf{x}}_{t,j}, \mathbf{S}_{t,j})$ .

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

For example, let  $(\mathbf{0}, \mathbf{I}_p)$  and  $(\mathbf{1}, \text{diag}(1, 3, \dots, p))$  be the high breakdown attractors where  $\mathbf{0}$  and  $\mathbf{1}$  are the  $p \times 1$  vectors of zeroes and ones. If the

minimum determinant criterion is used, then the final estimator is  $(\mathbf{0}, \mathbf{I}_p)$ . Although the MCD criterion is used, the algorithm estimator does not have the same properties as the MCD estimator.

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

This theory shows that the Maronna et al. (2006, pp. 198–199) estimators that use  $K = 500$  and one concentration step ( $k = 0$ ) are inconsistent and zero breakdown. The following theorem is useful because it does not depend on the criterion used to choose the attractor. If the algorithm needs to use many attractors to achieve outlier resistance, then the individual attractors have little outlier resistance. Such estimators include elemental concentration algorithms, heuristic and genetic algorithms, and projection algorithms. Algorithms where all  $K$  of the attractors are inconsistent, such as elemental concentration algorithms that use  $k$  concentration steps, are especially untrustworthy. As another example, Stahel–Donoho algorithms use randomly chosen projections and the attractor is a weighted mean and covariance matrix computed for each projection. If randomly chosen projections result in inconsistent attractors, then the Stahel–Donoho algorithm is likely inconsistent.

Suppose there are  $K$  consistent estimators  $(T_j, \mathbf{C}_j)$  of  $(\boldsymbol{\mu}, a \boldsymbol{\Sigma})$  for some constant  $a > 0$ , each with the same rate  $n^\delta$ . If  $(T_A, \mathbf{C}_A)$  is an estimator obtained by choosing one of the  $K$  estimators, then  $(T_A, \mathbf{C}_A)$  is a consistent estimator of  $(\boldsymbol{\mu}, a \boldsymbol{\Sigma})$  with rate  $n^\delta$  by Pratt (1959). See Theorem 3.16.

**Theorem 4.5.** Suppose the algorithm estimator chooses an attractor as the final estimator where there are  $K$  attractors and  $K$  is fixed.

- i) If all of the attractors are consistent estimators of  $(\boldsymbol{\mu}, a \boldsymbol{\Sigma})$ , then the algorithm estimator is a consistent estimator of  $(\boldsymbol{\mu}, a \boldsymbol{\Sigma})$ .
- ii) If all of the attractors are consistent estimators of  $(\boldsymbol{\mu}, a \boldsymbol{\Sigma})$  with the same rate, e.g.,  $n^\delta$  where  $0 < \delta \leq 0.5$ , then the algorithm estimator is a consistent estimator of  $(\boldsymbol{\mu}, a \boldsymbol{\Sigma})$  with the same rate as the attractors.
- iii) If all of the attractors are high breakdown, then the algorithm estimator is high breakdown.
- iv) Suppose the data  $\mathbf{x}_1, \dots, \mathbf{x}_n$  are iid and  $P(\mathbf{x}_i = \boldsymbol{\mu}) < 1$ . The elemental basic resampling algorithm estimator ( $k = -1$ ) is inconsistent.
- v) The elemental concentration algorithm is zero breakdown.



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

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

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

Since the FMCD estimator is a zero breakdown elemental concentration algorithm, the Hubert et al. (2008) claim that “MCD can be efficiently computed with the FAST-MCD estimator” is false. Suppose  $K$  is fixed, but at least one randomly drawn start is iterated to convergence so that  $k$  is not fixed. Then it is not known whether the attractors are inconsistent or consistent estimators, so it is not known whether FMCD is consistent. It is possible to produce consistent estimators if  $K \equiv K_n$  is allowed to increase to  $\infty$ .

**Remark 4.1.** Let  $\gamma_o$  be the highest percentage of large outliers that an elemental concentration algorithm can detect reliably. For many data sets,

$$\gamma_o \approx \min \left( \frac{n - c_n}{n}, 1 - [1 - (0.2)^{1/K}]^{1/h} \right) 100\% \quad (4.5)$$

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

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

$0.8 = 1 - [1 - (1 - \gamma)^h]^K$ . Fix the number of starts  $K$  and solve this equation for  $\gamma$ .  $\square$

Equation (4.5) agrees very well with the Rousseeuw and Van Driessen (1999) simulation performed on the hybrid FMCD algorithm that uses both concentration and partitioning. Section 4.4 will provide theory for the useful practical 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.

## 4.4 Theory for Practical Estimators

It is convenient to let the  $\mathbf{x}_i$  be random vectors for large sample theory, but the  $\mathbf{x}_i$  are fixed clean observed data vectors when discussing breakdown. This section presents the FCH estimator to be used along with the classical and FMCD estimators. Recall from Definition 4.5 that a *concentration algorithm* uses  $K_n$  starts  $(T_{-1,j}, \mathbf{C}_{-1,j})$ . After finding  $(T_{0,j}, \mathbf{C}_{0,j})$ , each start is refined with  $k$  concentration steps, resulting in  $K_n$  attractors  $(T_{k,j}, \mathbf{C}_{k,j})$ , and the concentration attractor  $(T_A, \mathbf{C}_A)$  is the attractor that optimizes the criterion.

Concentration algorithms include the *basic resampling algorithm* as a special case with  $k = -1$ . Using  $k = 10$  concentration steps works well, and iterating until convergence is usually fast. The DGK estimator (Devlin et al. 1975, 1981) defined below is one example. Gnanadesikan and Kettenring (1972, pp. 94–95) provided 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 4.1. This section will show that the Olive (2004a) MB estimator is a high breakdown estimator and that the DGK estimator is a  $\sqrt{n}$  consistent estimator of  $(\boldsymbol{\mu}, a_{MCD} \boldsymbol{\Sigma})$ , the same quantity estimated by the MCD estimator. Both estimators use the classical estimator computed from  $c_n \approx n/2$  cases. The breakdown point of the DGK estimator has been conjectured to be “at most  $1/p$ .” See Rousseeuw and Leroy (1987, p. 254). Gnanadesikan (1977, p. 134) provided an estimator somewhat similar to the MB estimator.

**Definition 4.6.** The *DGK estimator*  $(T_{k,D}, \mathbf{C}_{k,D}) = (T_{DGK}, \mathbf{C}_{DGK})$  uses the classical estimator  $(T_{-1,D}, \mathbf{C}_{-1,D}) = (\bar{\mathbf{x}}, \mathbf{S})$  as the only start.

**Definition 4.7.** The *median ball (MB) estimator*  $(T_{k,M}, \mathbf{C}_{k,M}) = (T_{MB}, \mathbf{C}_{MB})$  uses  $(T_{-1,M}, \mathbf{C}_{-1,M}) = (\text{MED}(\mathbf{W}), \mathbf{I}_p)$  as the only start where  $\text{MED}(\mathbf{W})$  is the coordinatewise median. So  $(T_{0,M}, \mathbf{C}_{0,M})$  is the classical estimator applied to the “half set” of data closest to  $\text{MED}(\mathbf{W})$  in Euclidean distance.

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

**Theorem 4.6.** Suppose  $(T, \mathbf{C})$  is a high breakdown estimator where  $\mathbf{C}$  is a symmetric, positive definite  $p \times p$  matrix if the contamination proportion  $d_n/n$  is less than the breakdown value. Then the concentration attractor  $(T_k, \mathbf{C}_k)$  is a high breakdown estimator if the coverage  $c_n \approx n/2$  and the data are in general position.

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

$$0 < \|\mathbf{x}\|^2 \tau_p \leq \mathbf{x}^T \mathbf{A} \mathbf{x} \leq \|\mathbf{x}\|^2 \tau_1. \quad (4.6)$$

Let  $\lambda_1 \geq \dots \geq \lambda_p$  be the eigenvalues of  $\mathbf{C}$ . By (4.6),

$$\frac{1}{\lambda_1} \|\mathbf{x} - T\|^2 \leq (\mathbf{x} - T)^T \mathbf{C}^{-1} (\mathbf{x} - T) \leq \frac{1}{\lambda_p} \|\mathbf{x} - T\|^2. \quad (4.7)$$

By (4.7), if the  $D_{(i)}^2$  are the order statistics of the  $D_i^2(T, \mathbf{C})$ , then  $D_{(i)}^2 < V$  for some constant  $V$  that depends on the clean data but not on the outliers even if  $i$  and  $d_n$  are near  $n/2$ . (Note that  $1/\lambda_p$  and  $\text{MED}(\|\mathbf{x}_i - T\|^2)$  are both bounded for high breakdown estimators even for  $d_n$  near  $n/2$ .)

Following Johnson and Wichern (1988, pp. 50, 103), the boundary of the set  $\{\mathbf{x} | D_{\mathbf{x}}^2 \leq h^2\} = \{\mathbf{x} | (\mathbf{x} - T)^T \mathbf{C}^{-1} (\mathbf{x} - T) \leq h^2\}$  is a hyperellipsoid centered at  $T$  with axes of length  $2h\sqrt{\lambda_i}$ . Hence  $\{\mathbf{x} | D_{\mathbf{x}}^2 \leq D_{(c_n)}^2\}$  is contained in some ball about the origin of radius  $r$  where  $r$  does not depend on the number of outliers even for  $d_n$  near  $n/2$ . This is the set containing the cases used to compute  $(T_0, \mathbf{C}_0)$ . Since the set is bounded,  $T_0$  is bounded and the largest eigenvalue  $\lambda_{1,0}$  of  $\mathbf{C}_0$  is bounded by Lemma 4.3. The determinant  $\det(\mathbf{C}_{MCD})$  of the HB minimum covariance determinant estimator satisfies  $0 < \det(\mathbf{C}_{MCD}) \leq \det(\mathbf{C}_0) = \lambda_{1,0} \dots \lambda_{p,0}$ , and  $\lambda_{p,0} > \inf \det(\mathbf{C}_{MCD}) / \lambda_{1,0}^{p-1} > 0$  where the infimum is over all possible data sets with  $n - d_n$  clean cases and  $d_n$  outliers. Since these bounds do not depend on the outliers even for  $d_n$  near  $n/2$ ,  $(T_0, \mathbf{C}_0)$  is a high breakdown estimator. Now repeat the argument with  $(T_0, \mathbf{C}_0)$  in place of  $(T, \mathbf{C})$  and  $(T_1, \mathbf{C}_1)$  in place of  $(T_0, \mathbf{C}_0)$ . Then  $(T_1, \mathbf{C}_1)$  is high breakdown. Repeating the argument iteratively shows  $(T_k, \mathbf{C}_k)$  is high breakdown.  $\square$

The following corollary shows that it is easy to find a subset  $J$  of  $c_n \approx n/2$  cases such that the classical estimator  $(\bar{\mathbf{x}}_J, \mathbf{S}_J)$  applied to  $J$  is a HB estimator of MLD.

**Corollary 4.7.** Let  $J$  consist of the  $c_n$  cases  $\mathbf{x}_i$  such that  $\|\mathbf{x}_i - \text{MED}(\mathbf{W})\| \leq \text{MED}(\|\mathbf{x}_i - \text{MED}(\mathbf{W})\|)$ . Then the classical estimator  $(\bar{\mathbf{x}}_J, \mathbf{S}_J)$  applied to  $J$  is a HB estimator of MLD.

To investigate the consistency and rate of robust estimators of multivariate location and dispersion, review Definitions 3.14 and 3.15.

The following assumption (E1) gives a class of distributions where we can prove that the new robust estimators are  $\sqrt{n}$  consistent. Cator and Lopuhaä (2010, 2012) showed that MCD is consistent provided that the MCD functional is unique. Distributions where the functional is unique are called “unimodal,” and rule out, for example, a spherically symmetric uniform distribution. Theorem 4.8 is crucial for theory, and Theorem 4.9 shows that under (E1), both MCD and DGK are estimating  $(\boldsymbol{\mu}, a_{\text{MCD}}\boldsymbol{\Sigma})$ .

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

Lopuhaä (1999) showed that if a start  $(T, \mathbf{C})$  is a consistent affine equivariant estimator of  $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ , then the classical estimator applied to the cases with  $D_i^2(T, \mathbf{C}) \leq h^2$  is a consistent estimator of  $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$  where  $a, s > 0$  are some constants. Affine equivariance is not used for  $\boldsymbol{\Sigma} = \mathbf{I}_p$ . Also, the attractor and the start have the same rate. If the start is inconsistent, then so is the attractor. The weight function  $I(D_i^2(T, \mathbf{C}) \leq h^2)$  is an indicator that is 1 if  $D_i^2(T, \mathbf{C}) \leq h^2$  and 0 otherwise.

**Theorem 4.8, Lopuhaä (1999).** Assume the number of concentration steps  $k$  is fixed. a) If the start  $(T, \mathbf{C})$  is inconsistent, then so is the attractor.

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

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

Let  $\delta = 0.5$ . Applying Theorem 4.8c) iteratively for a fixed number  $k$  of steps produces a sequence of estimators  $(T_0, \mathbf{C}_0), \dots, (T_k, \mathbf{C}_k)$  where  $(T_j, \mathbf{C}_j)$

is a  $\sqrt{n}$  consistent affine equivariant estimator of  $(\boldsymbol{\mu}, a_j \boldsymbol{\Sigma})$  where the constants  $a_j > 0$  depend on  $s, h, p$ , and the elliptically contoured distribution, but do not otherwise depend on the consistent start  $(T, \mathbf{C}) \equiv (T_{-1}, \mathbf{C}_{-1})$ .

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

**Conjecture 4.1.** Change the finite 4th moments assumption to a finite 2nd moments in assumption E1). Suppose  $(T, \mathbf{C})$  is a consistent estimator of  $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$  with rate  $n^\delta$  where  $s > 0$  and  $0 < \delta \leq 0.5$ . Then the classical estimator applied to the cases with  $D_i^2(T, \mathbf{C}) \leq h^2$  is a consistent estimator of  $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$  with the same rate  $n^\delta$  where  $a > 0$ .

**Remark 4.2.** To see that the Lopuhaä (1999) theory extends to concentration where the weight function uses  $h^2 = D_{(c_n)}^2(T, \mathbf{C})$ , note that  $(T, \tilde{\mathbf{C}}) \equiv (T, D_{(c_n)}^2(T, \mathbf{C}) \mathbf{C})$  is a consistent estimator of  $(\boldsymbol{\mu}, b\boldsymbol{\Sigma})$  where  $b > 0$  is derived in (4.9), and weight function  $I(D_i^2(T, \tilde{\mathbf{C}}) \leq 1)$  is equivalent to the concentration weight function  $I(D_i^2(T, \mathbf{C}) \leq D_{(c_n)}^2(T, \mathbf{C}))$ . As noted above, Proposition 4.1,  $(T, \tilde{\mathbf{C}})$  is affine equivariant if  $(T, \mathbf{C})$  is affine equivariant. Hence Lopuhaä (1999) theory applied to  $(T, \tilde{\mathbf{C}})$  with  $h = 1$  is equivalent to theory applied to affine equivariant  $(T, \mathbf{C})$  with  $h^2 = D_{(c_n)}^2(T, \mathbf{C})$ .

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

$$\begin{aligned} & (\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)^T [\mathbf{C}^{-1} - s^{-1} \boldsymbol{\Sigma}^{-1} + s^{-1} \boldsymbol{\Sigma}^{-1}] (\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T) \\ & = s^{-1} D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(n^{-\delta}). \end{aligned} \quad (4.8)$$

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

$$b = D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (4.9)$$

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

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

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

be the population median of the population squared distances. By Remark 4.2, if  $(T, \mathbf{C})$  is a  $\sqrt{n}$  consistent affine equivariant estimator of  $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$  then

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

Theorem 4.9 shows that  $a = a_{MCD}$  where  $\xi = 0.5$ . Hence concentration with a consistent affine equivariant estimator of  $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$  with rate  $n^\delta$  as a start results in a consistent affine equivariant estimator of  $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$  with rate  $n^\delta$ . This result can be applied iteratively for a finite number of concentration steps. Hence DGK is a  $\sqrt{n}$  consistent affine equivariant estimator of the same quantity that MCD is estimating. It is not known if the results hold if concentration is iterated to convergence. For multivariate normal data,  $D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \chi_p^2$ .

**Theorem 4.9.** Assume that (E1) holds and that  $(T, \mathbf{C})$  is a consistent affine equivariant estimator of  $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$  with rate  $n^\delta$  where the constants  $s > 0$  and  $0 < \delta \leq 0.5$ . Then the classical estimator  $(\bar{\mathbf{x}}_{t,j}, \mathbf{S}_{t,j})$  computed from the  $c_n \approx n/2$  of cases with the smallest distances  $D_i(T, \mathbf{C})$  is a consistent affine equivariant estimator of  $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$  with the same rate  $n^\delta$ .

**Proof.** By Remark 4.2, the estimator is a consistent affine equivariant estimator of  $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$  with rate  $n^\delta$ . By the remarks above,  $a$  will be the same for any consistent affine equivariant estimator of  $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$  and  $a$  does not depend on  $s > 0$  or  $\delta \in (0, 0.5]$ . Hence the result follows if  $a = a_{MCD}$ . The MCD estimator is a  $\sqrt{n}$  consistent affine equivariant estimator of  $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$  by Cator and Lopuhaä (2010, 2012). If the MCD estimator is the start, then it is also the attractor by Rousseeuw and Van Driessen (1999) who show that concentration does not increase the MCD criterion. Hence  $a = a_{MCD}$ .  $\square$

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

**Definition 4.8.** Let the “median ball” be the hypersphere containing the “half set” of data closest to  $\text{MED}(\mathbf{W})$  in Euclidean distance. The *FCH estimator* uses the MB attractor if the DGK location estimator  $T_{DGK}$  is outside of the median ball, and the attractor with the smallest determinant, otherwise. Let  $(T_A, \mathbf{C}_A)$  be the attractor used. Then the estimator  $(T_{FCH}, \mathbf{C}_{FCH})$  takes  $T_{FCH} = T_A$  and

$$\mathbf{C}_{FCH} = \frac{\text{MED}(D_i^2(T_A, \mathbf{C}_A))}{\chi_{p,0.5}^2} \mathbf{C}_A \quad (4.10)$$

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

**Remark 4.3.** The *MBA estimator*  $(T_{MBA}, \mathbf{C}_{MBA})$  uses the attractor  $(T_A, \mathbf{C}_A)$  with the smallest determinant. Hence the DGK estimator is used as the attractor if  $\det(\mathbf{C}_{DGK}) \leq \det(\mathbf{C}_{MB})$ , and the MB estimator is used as the attractor, otherwise. Then  $T_{MBA} = T_A$  and  $\mathbf{C}_{MBA}$  are computed using the right-hand side of (4.10). The difference between the FCH and MBA estimators is that the FCH estimator also uses a location criterion to choose the attractor: if the DGK location estimator  $T_{DGK}$  has a greater Euclidean distance from  $\text{MED}(\mathbf{W})$  than half the data, then FCH uses the MB attractor. The FCH estimator only uses the attractor with the smallest determinant if  $\|T_{DGK} - \text{MED}(\mathbf{W})\| \leq \text{MED}(D_i(\text{MED}(\mathbf{W}), \mathbf{I}_p))$ . Using the location criterion increases the outlier resistance of the FCH estimator for certain types of outliers, as will be seen in Section 4.5.

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

**Theorem 4.10.**  $T_{FCH}$  is high breakdown if the clean data are in general position. Suppose (E1) holds. If  $(T_A, \mathbf{C}_A)$  is the DGK or MB attractor with the smallest determinant, then  $(T_A, \mathbf{C}_A)$  is a  $\sqrt{n}$  consistent estimator of  $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ . Hence the MBA and FCH estimators are outlier resistant  $\sqrt{n}$  consistent estimators of  $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$  where  $c = u_{0.5}/\chi_{p,0.5}^2$  and  $c = 1$  for multivariate normal data.

**Proof.**  $T_{FCH}$  is high breakdown since it is a bounded distance from  $\text{MED}(\mathbf{W})$  even if the number of outliers is close to  $n/2$ . Under (E1), the FCH and MBA estimators are asymptotically equivalent since  $\|T_{DGK} - \text{MED}(\mathbf{W})\| \rightarrow 0$  in probability. The estimator satisfies  $0 < \det(\mathbf{C}_{MCD}) \leq \det(\mathbf{C}_A) \leq \det(\mathbf{C}_{0,M}) < \infty$  by Theorem 4.6 if up to nearly 50% of the cases are outliers. If the distribution is spherical about  $\boldsymbol{\mu}$ , then the result follows from Pratt (1959) and Proposition 4.4 since both starts are  $\sqrt{n}$  consistent. Otherwise, the MB estimator  $\mathbf{C}_{MB}$  is a biased estimator of  $a_{MCD}\boldsymbol{\Sigma}$ . But the DGK estimator  $\mathbf{C}_{DGK}$  is a  $\sqrt{n}$  consistent estimator of  $a_{MCD}\boldsymbol{\Sigma}$  by Theorem 4.9 and  $\|\mathbf{C}_{MCD} - \mathbf{C}_{DGK}\| = O_P(n^{-1/2})$ . Thus the probability that the DGK attractor minimizes the determinant goes to one as  $n \rightarrow \infty$ , and  $(T_A, \mathbf{C}_A)$  is asymptotically equivalent to the DGK estimator  $(T_{DGK}, \mathbf{C}_{DGK})$ .

Let  $\mathbf{C}_F = \mathbf{C}_{FCH}$  or  $\mathbf{C}_F = \mathbf{C}_{MBA}$ . Let  $P(U \leq u_\alpha) = \alpha$  where  $U$  is given by (3.9). Then the scaling in (4.10) makes  $\mathbf{C}_F$  a consistent estimator of  $c\boldsymbol{\Sigma}$  where  $c = u_{0.5}/\chi_{p,0.5}^2$ , and  $c = 1$  for multivariate normal data.  $\square$

Many variants of the FCH and MBA estimators can be given where the algorithm gives a  $\sqrt{n}$  consistent estimator of  $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ . One such variant uses  $K$  starts  $(T_{-1,j}, \mathbf{C}_{-1,j})$  that are affine equivariant  $\sqrt{n}$  consistent estimators of  $(\boldsymbol{\mu}, s_j \boldsymbol{\Sigma})$  where  $s_j > 0$ . The MCD criterion is used to choose the final attractor, and scaling is done as in (4.10). A second variant is the same as the first, but the  $K$ th attractor is replaced by the MB estimator, and for  $j < K$  the  $j$ th attractor  $(T_{k,j}, \mathbf{C}_{k,j})$  is not used if  $T_{k,j}$  has a greater Euclidean distance from  $\text{MED}(\mathbf{X})$  than half the data. Then the location estimator of the algorithm is high breakdown.

Suppose the attractor is  $(\bar{\mathbf{x}}_{k,j}, \mathbf{S}_{k,j})$  computed from a subset of  $c_n$  cases. The  $\text{MCD}(c_n)$  criterion is the determinant  $\det(\mathbf{S}_{k,j})$ . The volume of the hyperellipsoid  $\{\mathbf{z} : (\mathbf{z} - \bar{\mathbf{x}}_{k,j})^T \mathbf{S}_{k,j}^{-1} (\mathbf{z} - \bar{\mathbf{x}}_{k,j}) \leq h^2\}$  is equal to

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p \sqrt{\det(\mathbf{S}_{k,j})}, \quad (4.11)$$

see Johnson and Wichern (1988, pp. 103–104). The “MVE( $c_n$ )” criterion is  $h^p \sqrt{\det(\mathbf{S}_{k,j})}$  where  $h = D_{(c_n)}(\bar{\mathbf{x}}_{k,j}, \mathbf{S}_{k,j})$  (but does not actually correspond to the minimum volume ellipsoid (MVE) estimator).

We also considered several estimators that use the MB and DGK estimators as attractors. CMVE is a concentration algorithm like FCH, but the “MVE” criterion is used in place of the MCD criterion. A standard method of reweighting can be used to produce the RMBA, RFCH, and RCMVE estimators. RMVN uses a slightly modified method of reweighting so that RMVN gives good estimates of  $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  for multivariate normal data, even when certain types of outliers are present.

**Definition 4.9.** The *RFCH estimator* uses two standard reweighting steps. Let  $(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1)$  be the classical estimator applied to the  $n_1$  cases with  $D_i^2(T_{FCH}, \mathbf{C}_{FCH}) \leq \chi_{p,0.975}^2$ , and let

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

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

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

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



**Definition 4.10.** The *RMVN estimator* uses  $(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1)$  and  $n_1$  as above. Let  $q_1 = \min\{0.5(0.975)n/n_1, 0.995\}$ , and

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

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

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

**Definition 4.11.** Let the  $n_2$  cases in Definition 4.10 be known as the *RMVN set*  $U$ . Hence  $(T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2) = (\bar{\boldsymbol{x}}_U, \boldsymbol{S}_U)$  is the classical estimator applied to the RMVN set  $U$ , which can be regarded as the untrimmed data (the data not trimmed by ellipsoidal trimming) or the cleaned data. Also  $\boldsymbol{S}_U$  is the unscaled estimated dispersion matrix while  $\boldsymbol{C}_{RMVN}$  is the scaled estimated dispersion matrix.

**Remark 4.4.** Classical methods will be applied to the RMVN subset  $U$  to make robust methods throughout this text. Under (E1),  $(\bar{\boldsymbol{x}}_U, \boldsymbol{S}_U)$  is a  $\sqrt{n}$  consistent estimator of  $(\boldsymbol{\mu}, c_U \boldsymbol{\Sigma})$  for some constant  $c_U > 0$  that depends on the underlying distribution of the iid  $\boldsymbol{x}_i$ . For a general estimator of multivariate location and dispersion  $(T_A, \boldsymbol{C}_A)$ , typically a reweight for efficiency step is performed, resulting in a set  $U$  such that the classical estimator  $(\bar{\boldsymbol{x}}_U, \boldsymbol{S}_U)$  is the classical estimator applied to a set  $U$ . For example, use  $U = \{\boldsymbol{x}_i | D_i^2(T_A, \boldsymbol{C}_A) \leq \chi_{p, 0.975}^2\}$ . Then the final estimator is  $(T_F, \boldsymbol{C}_F) = (\bar{\boldsymbol{x}}_U, a\boldsymbol{S}_U)$  where scaling is done as in Equation (4.10) in an attempt to make  $\boldsymbol{C}_F$  a good estimator of  $\boldsymbol{\Sigma}$  if the iid data are from a  $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  distribution. Then  $(\bar{\boldsymbol{x}}_U, \boldsymbol{S}_U)$  can be shown to be a  $\sqrt{n}$  consistent estimator of  $(\boldsymbol{\mu}, c_U \boldsymbol{\Sigma})$  for a large class of distributions for the RMVN set, for the RFCH set, or if  $(T_A, \boldsymbol{C}_A)$  is an affine equivariant  $\sqrt{n}$  consistent estimator of  $(\boldsymbol{\mu}, c_A \boldsymbol{\Sigma})$  on a large class of distributions. The necessary theory is not yet available for other practical robust reweighted estimators such as OGK and Det-MCD.

**Table 4.1** Average Dispersion Matrices for Near Point Mass Outliers

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

**Table 4.2** Average Dispersion Matrices for Mean Shift Outliers

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

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

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

Simulations suggested  $(T_{RMVN}, \mathbf{C}_{RMVN})$  gives useful estimates of  $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  for a variety of outlier configurations. Using 20 runs and  $n = 1000$ , the averages of the dispersion matrices were computed when the bulk of the data are iid  $N_2(\mathbf{0}, \boldsymbol{\Sigma})$  where  $\boldsymbol{\Sigma} = \text{diag}(1, 2)$ . For clean data, FCH, RFCH, and RMVN give  $\sqrt{n}$  consistent estimators of  $\boldsymbol{\Sigma}$ , while FMCD and the Maronna and Zamar (2002) OGK estimator seem to be approximately unbiased for  $\boldsymbol{\Sigma}$ . The median ball estimator was scaled using (4.10) and estimated  $\text{diag}(1.13, 1.85)$ .

Next the data had  $\gamma = 0.4$  and the outliers had  $\mathbf{x} \sim N_2((0, 15)^T, 0.0001\mathbf{I}_2)$ , a near point mass at the major axis. FCH, MB, and RFCH estimated  $2.6\boldsymbol{\Sigma}$  while RMVN estimated  $\boldsymbol{\Sigma}$ . FMCD and OGK failed to estimate  $d\boldsymbol{\Sigma}$ . Note that  $\chi_{2,5/6}^2/\chi_{2,0.5}^2 = 2.585$ . See Table 4.1. The following *R* commands were used where *mldsim* is from *mpack*.

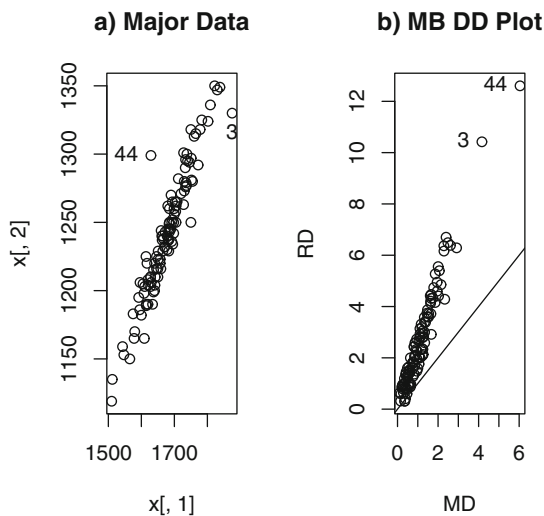
```
qchisq(5/6, 2)/qchisq(.5, 2) = 2.584963
mldsim(n=1000, p=2, outliers=6, pm=15)
```

Next the data had  $\gamma = 0.4$  and the outliers had  $\mathbf{x} \sim N_2((20, 20)^T, \boldsymbol{\Sigma})$ , a mean shift with the same covariance matrix as the clean cases. Rocke and Woodruff (1996) suggest that outliers with mean shift are hard to detect. FCH, FMCD, MB, and RFCH estimated  $2.6\boldsymbol{\Sigma}$  while RMVN estimated  $\boldsymbol{\Sigma}$ , and OGK failed. See Table 4.2. The *R* command is shown below.

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

**Remark 4.5.** The RFCH and RMVN estimators are recommended. If these estimators are too slow and outlier detection is of interest, try the RMB estimator, the reweighted MB estimator. If RMB is too slow or if  $n < 2(p + 1)$ , the Euclidean distances  $D_i(\text{MED}(\mathbf{W}), \mathbf{I})$  of  $\mathbf{x}_i$  from the coordinatewise median  $\text{MED}(\mathbf{W})$  may be useful. A DD plot of  $D_i(\bar{\mathbf{x}}, \mathbf{I})$  versus  $D_i(\text{MED}(\mathbf{W}), \mathbf{I})$  is also useful for outlier detection and for whether  $\bar{\mathbf{x}}$  and  $\text{MED}(\mathbf{W})$  are giving similar estimates of multivariate location. See Section 4.7.

**Example 4.1.** Tremearne (1911) recorded  $\text{height} = \mathbf{x}[1]$  and  $\text{height while kneeling} = \mathbf{x}[2]$  of 112 people. Figure 4.1a shows a scatterplot of the data. Case 3 has the largest Euclidean distance of 214.767 from  $\text{MED}(\mathbf{W}) = (1680, 1240)^T$ , but if the distances correspond to the contours of a covering ellipsoid, then case 44 has the largest distance. For  $k = 0$ ,  $(T_{0,M}, \mathbf{C}_{0,M})$  is the classical estimator applied to the “half set” of cases closest to  $\text{MED}(\mathbf{W})$

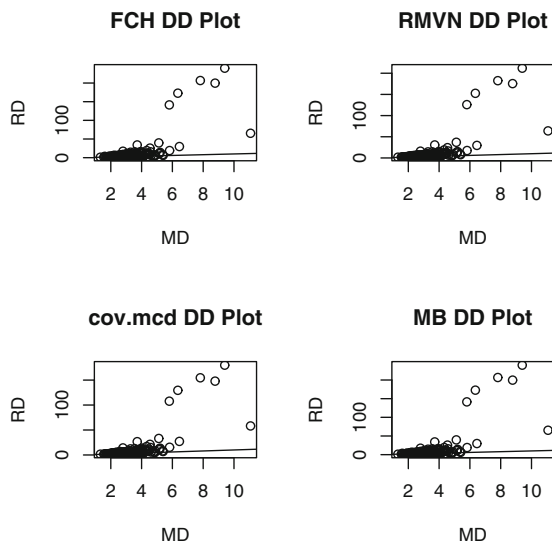


**Fig. 4.1** Plots for Major Data

in Euclidean distance. The hypersphere (circle) centered at  $\text{MED}(\mathbf{W})$  that covers half the data is small because the data density is high near  $\text{MED}(\mathbf{W})$ . 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 4.1b shows the DD plot of the classical distances versus the MB distances. Notice that both the classical and MB estimators give the largest distances to cases 3 and 44. Notice that case 44 could not be detected using marginal methods.

As the dimension  $p$  gets larger, outliers that cannot be detected by marginal methods (case 44 in Example 4.1) become harder to detect. When  $p = 3$  imagine that the clean data is a baseball bat or stick with one end at the SW corner of the bottom of the box (corresponding to the coordinate axes) and one end at the NE corner of the top of the box. If the outliers are a ball, there is much more room to hide them in the box than in a covering rectangle when  $p = 2$ .

**Example 4.2.** The estimators can be useful when the data is not elliptically contoured. The Gladstone (1905) data has 11 variables on 267 persons after death. Head measurements were *breadth*, *circumference*, *head height*, *length*, and *size* as well as *cephalic index* and *brain weight*. Age, height, and two categorical variables *ageclass* (0: under 20, 1: 20–45, 2: over 45) and *sex* were also given. Figure 4.2 shows the DD plots for the FCH, RMVN, `cov.mcd`, and MB estimators. The DD plots from the DGK, MBA, CMVE,



**Fig. 4.2** DD Plots for Gladstone Data

RCMVE, and RFCH estimators were similar, and the six outliers in Figure 4.2 correspond to the six infants in the data set.

Chapter 5 shows that if a consistent robust estimator is scaled as in (4.10), then the plotted points in the DD plot will cluster about the identity line with unit slope and zero intercept if the data is multivariate normal, and about some other line through the origin if the data is from some other elliptically contoured distribution with a nonsingular covariance matrix. Since

multivariate procedures tend to perform well for elliptically contoured data, the DD plot is useful even if outliers are not present.

## 4.5 Outlier Resistance and Simulations

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

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

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

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

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

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

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

**Table 4.3** Scaled Variance  $nS^2(T_p)$  and  $nS^2(C_{p,p})$ 

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

Table 4.3 considers  $W = T_p$  and  $W = C_{p,p}$  for eight estimators,  $p = 5$  and 10, and  $n = 10p$  and 5000, when  $\mathbf{x} \sim N_p(\mathbf{0}, \text{diag}(1, \dots, p))$ . For the classical estimator, denoted by CLAS,  $T_p = \bar{x}_p \sim N(0, p/n)$ , and  $nS^2(T_p) \approx p$  while  $C_{p,p}$  is the sample variance of  $n$  iid  $N(0, p)$  random variables. Hence  $nS^2(C_{p,p}) \approx 2p^2$ . RFCH, RMVN, FMCD, and OGK use a “reweight for efficiency” concentration step that uses a random number of cases with percentage close to 97.5%. These four estimators had similar behavior. DGK, FCH, and MB used about 50% of the cases and had similar behavior. By Lopuhaä (1999), estimators with less than  $\sqrt{n}$  rate still have zero efficiency after the reweighting. Although FMCD, MB, and OGK have not been proven to be  $\sqrt{n}$  consistent, their values did not blow up even for  $n = 5000$ .

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

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

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

The clean cases had  $\mathbf{x} \sim N_p(\mathbf{0}, \text{diag}(1, 2, \dots, p))$ . Outlier types were the mean shift  $\mathbf{x} \sim N_p(pm\mathbf{1}, \text{diag}(1, 2, \dots, p))$  where  $\mathbf{1} = (1, \dots, 1)^T$  and  $\mathbf{x} \sim N_p((0, \dots, 0, pm)^T, 0.0001\mathbf{I}_p)$ , a near point mass at the major axis. Notice that the clean data can be transformed to a  $N_p(\mathbf{0}, \mathbf{I}_p)$  distribution by multiplying  $\mathbf{x}_i$  by  $\text{diag}(1, 1/\sqrt{2}, \dots, 1/\sqrt{p})$ , and this transformation changes the location of the near point mass to  $(0, \dots, 0, pm/\sqrt{p})^T$ .

For near point mass outliers, a hyperellipsoid with very small volume can cover half of the data if the outliers are at one end of the hyperellipsoid and some of the clean data are at the other end. This half set will produce a classical estimator with very small determinant by (4.11). In the simulations for large  $\gamma$ , as the near point mass is moved very far away from the bulk of the data, only the classical, MB, and OGK estimators did not have numerical

**Table 4.4** Number of Times Mean Shift Outliers had the Largest Distances

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

difficulties. Since the MCD estimator has smaller determinant than DGK while MVE has smaller volume than DGK, estimators like FMCD and MBA

that use the MVE or MCD criterion without using location information will be vulnerable to these outliers. FMCD is also vulnerable to outliers if  $\gamma$  is slightly larger than  $\gamma_o$  given by (4.5).

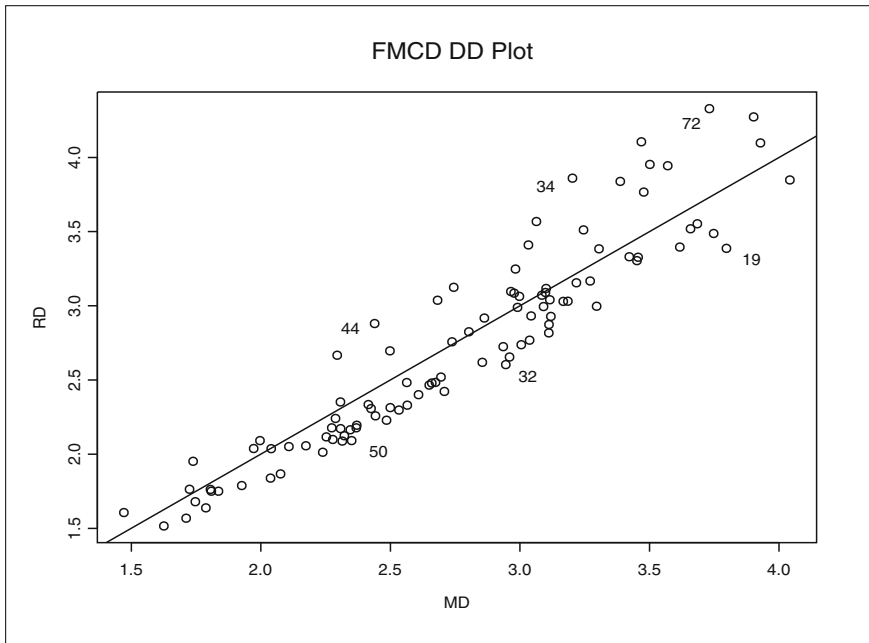
**Table 4.5** Number of Times Near Point Mass Outliers had the Largest Distances

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

Tables 4.4 and 4.5 help illustrate the results for the simulation. Large counts and small  $pm$  for fixed  $\gamma$  suggest greater ability to detect outliers. Values of  $p$  were 5, 10, 15, ..., 60. First consider the mean shift outliers and Table 4.4. For  $\gamma = 0.25$  and 0.4, MB usually had the highest counts. For  $5 \leq p \leq 20$  and the mean shift, the OGK estimator often had the smallest counts, and FMCD could not handle 40% outliers for  $p = 20$ . For  $25 \leq p \leq 60$ , OGK usually had the highest counts for  $\gamma = 0.05$  and 0.1. For  $p \geq 30$ , FMCD could not handle 25% outliers even for enormous values of  $pm$ .

In Table 4.5, FCH greatly outperformed MBA although the only difference between the two estimators is that FCH uses a location criterion as well as the MCD criterion. OGK performed well for  $\gamma = 0.05$  and  $20 \leq p \leq 60$  (not tabled). For large  $\gamma$ , OGK often has large bias for  $c\mathbf{\Sigma}$ . Then the outliers may need to be enormous before OGK can detect them. Also see Table 4.2, where OGK gave the outliers the largest distances for all runs, but  $\mathbf{C}_{OGK}$  does not give a good estimate of  $c\mathbf{\Sigma} = c \text{diag}(1, 2)$ .



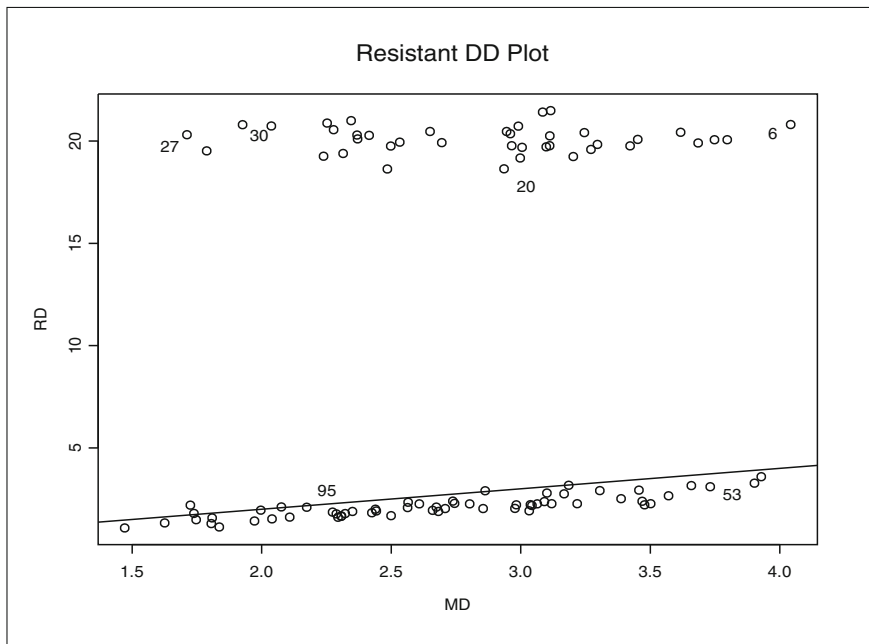


**Fig. 4.3** The FMCD Estimator Failed

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

In an older mean shift simulation, when  $p$  was 8 or larger, the `cov.mcd` estimator was usually not useful for detecting the mean shift outliers. Figure 4.3 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 4.4.

For many data sets, Equation (4.5) gives a rough approximation for the number of large outliers that concentration algorithms using  $K$  starts each consisting of  $h$  cases can handle. However, if the data set is multivariate and the bulk of the data falls in one compact hyperellipsoid while the outliers fall in another hugely distant compact hyperellipsoid, then a concentration algorithm using a single start can sometimes tolerate nearly 25% outliers.



**Fig. 4.4** The Outliers are Large in the MBA DD Plot

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  $N_p(\mathbf{0}, \mathbf{I}_p)$  distribution and the outliers come from the  $N_p(2000 \mathbf{1}, \mathbf{I}_p)$  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 4.11.** Consider the concentration and MCD estimators that both cover  $c_n$  cases. For multivariate data, if at least one of the starts is nonsingular, then the concentration attractor  $\mathbf{C}_A$  is less likely to be singular than the high breakdown MCD estimator  $\mathbf{C}_{MCD}$ .

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

### Software

The `robustbase` library was downloaded from ([www.r-project.org/#doc](http://www.r-project.org/#doc)). § 15.2 explains how to use the source command to get the `mpack` functions in *R* and how to download a library from *R*. Type the commands `library(MASS)` and `library(robustbase)` to compute the FMCD and OGK estimators with the `cov.mcd` and `covOGK` functions. To use Det-MCD instead of FMCD, change

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

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

The `mpack` function

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

can be used to produce Tables 4.1–4.5. Change outliers to 0 to examine the average of  $\hat{\boldsymbol{\mu}}$  and  $\hat{\boldsymbol{\Sigma}}$ . The function `mldsims6` is similar but does not need the `library` command since it compares the FCH, RFCH, CMVE, RCMVE, MB estimators, and the `covmb2` estimator of Section 4.7. The command 

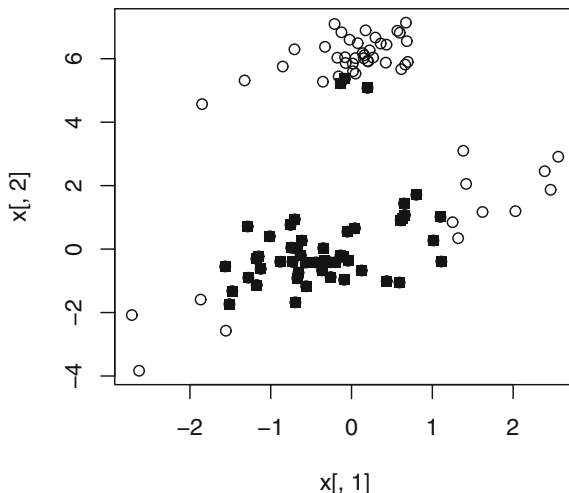
```
sctplt(n=200,p=10,gam=.2,outliers=3, pm=5)
```

will make an outlier data set. Then the FCH and MB DD plots are made (click on the right mouse button and highlight stop to go to the next plot) and then the scatterplot matrix. The scatterplot matrix can be used to determine whether the outliers are hard to detect with bivariate or univariate methods. If  $p > 10$  the bivariate plots may be too small. See Zhang (2011) for more simulations.

The function `covsim2` can be modified to show that the *R* implementation of FCH is usually much faster than OGK which is much faster than FMCD. The function `corrsim` can be used to simulate the correlations of robust distances with classical distances. 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. See Zhang (2011) for simulations.

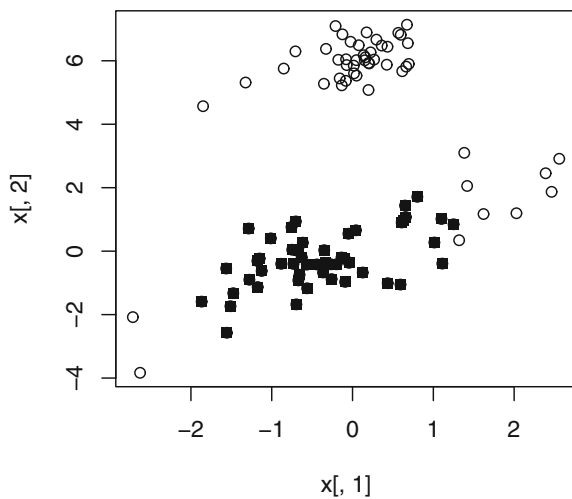


**Fig. 4.5** highlighted cases = half set with smallest RD =  $(T_0, C_0)$

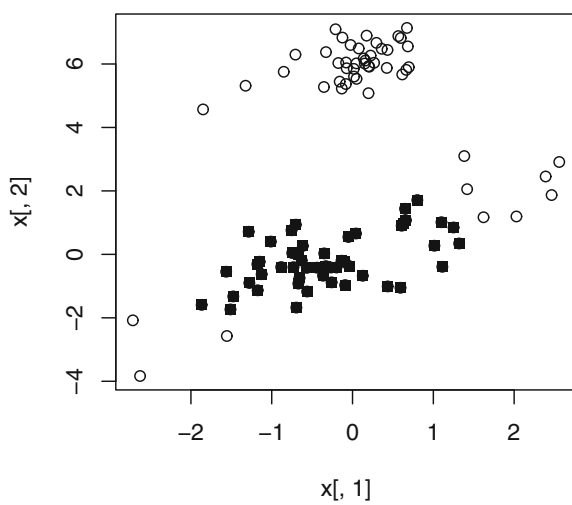
The function `cmve` computes CMVE and RCMVE, function `covfch` computes FCH and RFCH, while `covrmvn` computes the RMVN and MB estimators. The function `covrmmb` computes MB and RMB where RMB is like RMVN except the MB estimator is reweighted instead of FCH. Functions `covdggk`, `covmba`, and `rmmba`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

The `concmv` function described in Problem 4.5 illustrates concentration where the start is  $(\text{MED}(\mathbf{W}), \text{diag}([MAD(X_i)]^2))$ . In Figures 4.5, 4.6, and 4.7, the highlighted cases are the half set with the smallest distances, and the initial half set shown in Figure 4.5 is not clean, where  $n = 100$  and there are 40 outliers. The attractor shown in Figure 4.7 is clean. This type of data set has too many outliers for DGK while the MB starts and attractors are almost always clean.

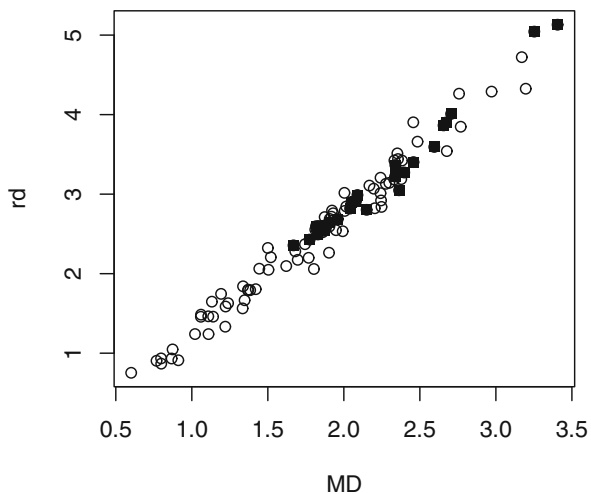
The `ddmv` function in Problem 4.6 illustrates concentration for the DGK estimator where the start is the classical estimator. Now  $n = 100, p = 4$ , and there are 25 outliers. A DD plot of classical distances MD versus robust distances RD is shown. See Figures 4.8, 4.9, 4.10, and 4.11. The half set of



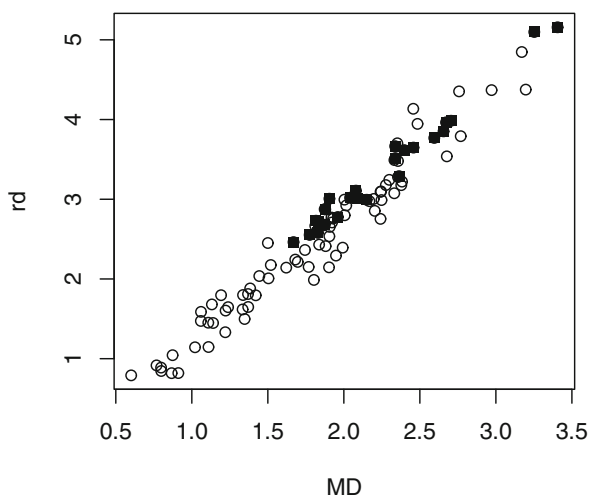
**Fig. 4.6** highlighted cases = half set with smallest RD =  $(T_1, C_1)$



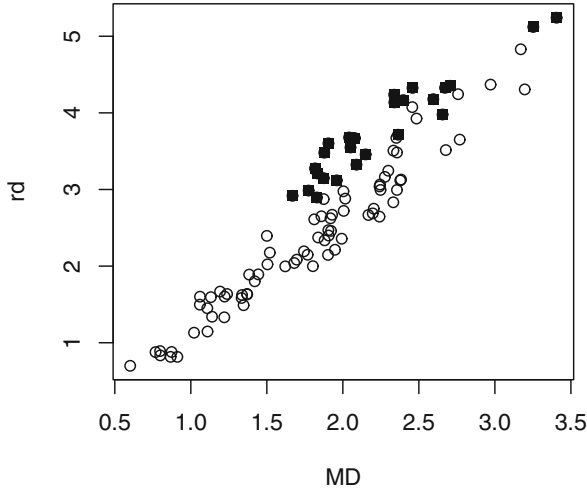
**Fig. 4.7** highlighted cases = half set with smallest RD =  $(T_2, C_2)$



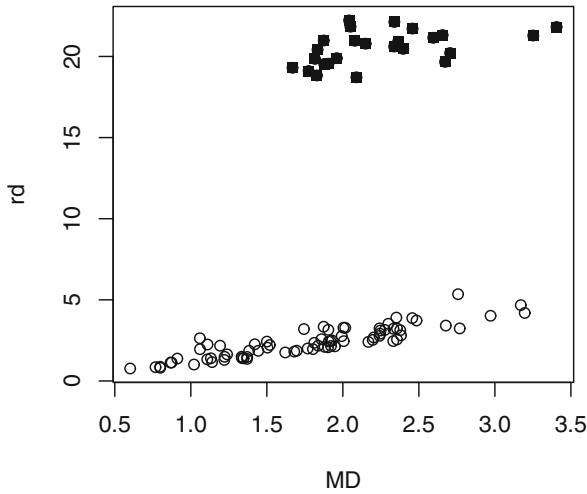
**Fig. 4.8** highlighted cases = outliers,  $RD = (T_{0,D}, C_{0,D})$



**Fig. 4.9** highlighted cases = outliers,  $RD = (T_{1,D}, C_{1,D})$



**Fig. 4.10** highlighted cases = outliers,  $RD = (T_{2,D}, C_{2,D})$



**Fig. 4.11** highlighted cases = outliers,  $RD = (T_{3,D}, C_{3,D})$

cases with the smallest RDs is used, and the initial half set shown in Figure 4.8 is not clean. The attractor in Figure 4.11 is the DGK estimator which uses a clean half set. The clean cases  $\mathbf{x}_i \sim N_4(\mathbf{0}, \text{diag}(1, 2, 3, 4))$  while the outliers  $\mathbf{x}_i \sim N_4((10, 10\sqrt{2}, 10\sqrt{3}, 20)^T, \text{diag}(1, 2, 3, 4))$ .

## 4.6 The RMVN and RFCH Sets

Both the RMVN and RFCH estimators compute the classical estimator  $(\bar{\mathbf{x}}_U, \mathbf{S}_U)$  on some set  $U$  containing  $n_U \geq n/2$  of the cases. Referring to Definition 4.9, for the RFCH estimator,  $(\bar{\mathbf{x}}_U, \mathbf{S}_U) = (T_{RFCH}, \tilde{\mathbf{S}}_2)$ , and then  $\mathbf{S}_U$  is scaled to form  $\mathbf{C}_{RFCH}$ . Referring to Definition 4.10, for the RMVN estimator,  $(\bar{\mathbf{x}}_U, \mathbf{S}_U) = (T_{RMVN}, \tilde{\mathbf{S}}_2)$ , and then  $\mathbf{S}_U$  is scaled to form  $\mathbf{C}_{RMVN}$ . See Definition 4.11.

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

Using the RMVN or RFCH, set  $U$  is simultaneously a plug in method and an objective way to clean the data such that the resulting robust method is often backed by theory. This result is extremely useful computationally: find the RMVN set or RFCH set  $U$ , then apply the classical method to the cases in the set  $U$ . This procedure is often equivalent to using  $(\bar{\mathbf{x}}_U, \mathbf{S}_U)$  as plug in estimators. The method can be applied if  $n > 2(p+1)$  but may not work well unless  $n > 20p$ . The *mpack* function `getu` gets the RMVN set  $U$  as well as the case numbers corresponding to the cases in  $U$ .

The set  $U$  is a small volume hyperellipsoid containing at least half of the cases since concentration is used. The set  $U$  can also be regarded as the “untrimmed data”: the data that was not trimmed by ellipsoidal trimming. Theory has been proved for a large class of elliptically contoured distributions, but it is conjectured that theory holds for a much wider class of distributions. See Conjectures 4.2 and 4.3 in Section 4.9. In simulations, RFCH and RMVN seem to estimate  $c\mathbf{\Sigma}\mathbf{x}$  if  $\mathbf{x} = \mathbf{A}\mathbf{z} + \boldsymbol{\mu}$  where  $\mathbf{z} = (z_1, \dots, z_p)^T$  and the  $z_i$  are iid from a continuous distribution with variance  $\sigma^2$ . Here  $\mathbf{\Sigma}\mathbf{x} = \text{Cov}(\mathbf{x}) = \sigma^2\mathbf{A}\mathbf{A}^T$ . The bias for the MB estimator seemed to be small. It is known that affine equivariant estimators give unbiased estimators of  $c\mathbf{\Sigma}\mathbf{x}$  if the distribution of  $z_i$  is also symmetric. DGK is affine equivariant, and RFCH and RMVN are asymptotically equivalent to a scaled DGK estimator. But in the simulations the results also held for skewed distributions.

In this text, usually the RMVN set  $U$  will be used. Several illustrative applications are given next, where the theory usually assumes that the cases are iid from a large class of elliptically contoured distributions. There are many other “robust methods” in the literature that use plug in estimators like FMCD. Replacing the plug in estimator by RMVN or RFCH will often greatly improve the robust method. See Chapter 14, and in Section 5.2, note that the prediction regions using RMVN performed much like the prediction regions using  $(\bar{\mathbf{x}}, \mathbf{S})$ .



i) The classical estimator of multivariate location and dispersion applied to the cases in  $U$  gives  $(\bar{\mathbf{x}}_U, \mathbf{S}_U)$ , a  $\sqrt{n}$  consistent estimator of  $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$  for some constant  $c > 0$ . See Remark 4.4.

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

iii) For principal component analysis (PCA), RPCA is the classical PCA method applied to the set  $U$ . See Theorems 6.1, 6.2, 6.3, and Section 6.2.

iv) For canonical correlation analysis (CCA), RCCA is the classical CCA method applied to the set  $U$ . See Theorem 7.1 and Section 7.2.

v) Let  $U_i$  be the RMVN or RFCH subset applied to the  $n_i$  cases from group  $i$  for  $i = 1, \dots, G$ . Let  $(\bar{\mathbf{x}}_{U_i}, \mathbf{S}_{U_i})$  be the sample mean and covariance applied to the cases in  $U_i$ . Let  $Y = i$  for cases in  $U_i$  which are from group  $i$ . Let  $U_{big} = U_1 \cup U_2 \cup \dots \cup U_G$  be the combined sample. Then apply the discriminant analysis method to  $U_{big}$  with the corresponding labels  $Y$ . For example, RFDA consists of applying classical FDA on  $U_{big}$ . See Section 8.9.

vi) For factor analysis, apply the factor analysis method to the set  $U$ . This method can be used as a diagnostic for methods such as the maximum likelihood method of factor analysis, but is backed by theory for principal component factor analysis. See Section 11.2.

vii) For multiple linear regression, let  $Y$  be the response variable,  $x_1 = 1$  and  $x_2, \dots, x_p$  be the predictor variables. Let  $\mathbf{z}_i = (Y_i, x_{i2}, \dots, x_{ip})^T$ . Let  $U$  be the RMVN or RFCH set formed using the  $\mathbf{z}_i$ . Then a classical regression estimator applied to the set  $U$  results in a robust regression estimator. For least squares, this is implemented with the *mpack* function `rmreg2`.

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

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

Let  $U$  be the RMVN or RFCH set formed using the  $\mathbf{z}_i$ . Then a classical least squares multivariate linear regression estimator applied to the set  $U$  results in a robust multivariate linear regression estimator. For least squares, this is implemented with the *mpack* function `rmreg2`. The method for multiple linear regression in vii) corresponds to  $m = 1$ . See Section 12.6.2.

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

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

## 4.7 What if $p > n$ ?

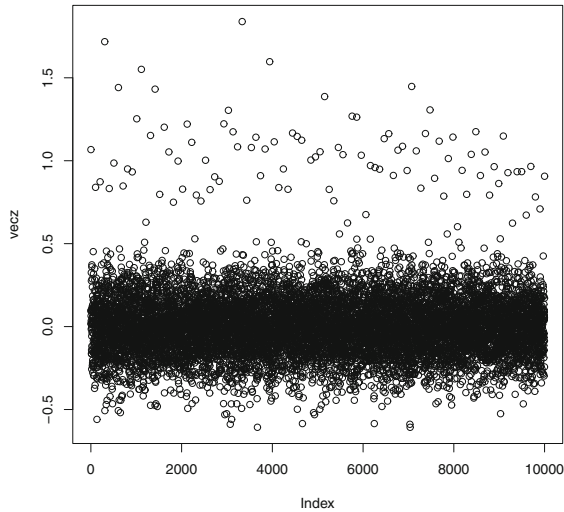
Most of the methods in this text work best if  $n \geq 20p$ , but often data sets have  $p > n$ , and outliers are a major problem. Usually the population covariance matrix is singular when  $p > n$ , but the sample covariance matrix, MB estimator with no concentration steps, and the sign covariance matrix can be computed. Weighted versions of the last two estimators are useful since concentration steps need nonsingular dispersion matrices. Compute the squared Euclidean distances of the  $\mathbf{x}_i$  from the coordinatewise median  $D_i^2 = D_i^2(\text{MED}(\mathbf{W}), \mathbf{I}_p)$ . Concentration type steps compute the weighted median  $\text{MED}_j$ : the coordinatewise median computed from the cases  $\mathbf{x}_i$  with  $D_i^2 \leq \text{MED}(D_i^2(\text{MED}_{j-1}, \mathbf{I}_p))$  where  $\text{MED}_0 = \text{MED}(\mathbf{W})$ . We often used  $j = 0$  (no concentration type steps) or  $j = 9$ . Let  $D_i = D_i(\text{MED}_j, \mathbf{I}_p)$ . Let  $W_i = 1$  if  $D_i \leq \text{MED}(D_1, \dots, D_n) + k\text{MAD}(D_1, \dots, D_n)$  where  $k \geq 0$  and  $k = 5$  is the default choice. Let  $W_i = 0$ , otherwise. This weighting corresponds to the weighting that would be used in a one sided metrically trimmed mean (Huber type skipped mean) of the distances. Just as classical methods can be applied to the set RMVN set  $U$  to create a robust estimator, classical methods for the  $p > n$  case, such as PLS (partial least squares), can be applied to the `covmb2` set  $B$  defined below.

**Definition 4.12.** Let the *covmb2* set  $B$  of at least  $n/2$  cases correspond to the cases with weight 1. Then the *covmb2* estimator  $(T, \mathbf{C})$  is the sample mean and sample covariance matrix applied to the cases in set  $B$ . Hence

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

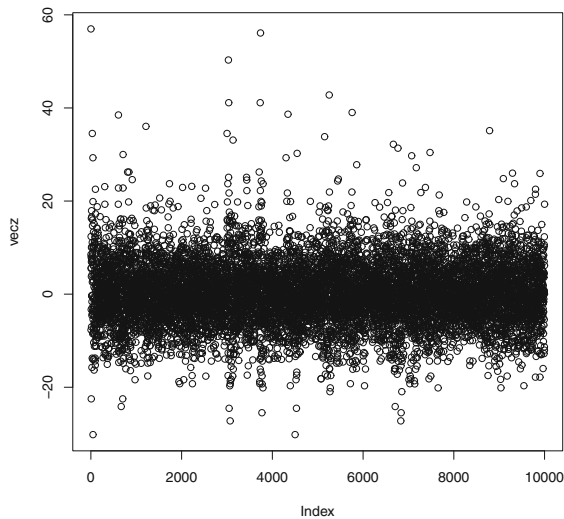
The `covmb2` estimator attempts to give a robust dispersion estimator that reduces the bias by using a big ball about the coordinatewise median instead of a ball that contains half of the cases. The weighting is the default method, but you can also plot the squared Euclidean distances and estimate the number  $m \geq n/2$  of cases with the smallest distances to be used. The *mpack* function `medout` makes the plot, and the *mpack* function `getB` gives the set  $B$  of cases that got weight 1 along with the index `indx` of the case numbers that got weight 1. The function `vecw` stacks the columns of the dispersion matrix  $\mathbf{C}$  into a vector. Then the elements of the matrix can be plotted.

The function `ddplot5` plots the Euclidean distances from the coordinatewise median versus the Euclidean distances from the `covmb2` location estimator. Typically, the plotted points in this DD plot cluster about the identity line, and outliers appear in the upper right corner of the plot with a gap between the bulk of the data and the outliers. An alternative for outlier detection is to replace  $\mathbf{C}$  by  $\mathbf{C}_d = \text{diag}(\hat{\sigma}_{11}, \dots, \hat{\sigma}_{pp})$ . For example, use  $\hat{\sigma}_{ii} = \mathbf{C}_{ii}$ . See Ro et al. (2015) and Tarr et al. (2016) for references.



**Fig. 4.12** Elements of  $\mathbf{C}$  for outlier data

**Example 4.3.** This example helps illustrate the effect of outliers on classical methods. The artificial data set had  $n = 50, p = 100$ , and the clean data was iid  $N_p(\mathbf{0}, \mathbf{I}_p)$ . Hence the diagonal elements of the population covariance matrix are 0, and the diagonal elements are 1. Plots of the elements of the sample covariance matrix  $\mathbf{S}$  and the `covmb2` estimator  $\mathbf{C}$  are not shown, but were similar to Figure 4.12. Then the first ten cases were contaminated:  $\mathbf{x}_i \sim N_p(\boldsymbol{\mu}, 100\mathbf{I}_p)$  where  $\boldsymbol{\mu} = (10, 0, \dots, 0)^T$ . Figure 4.12 shows that the `covmb2` dispersion matrix  $\mathbf{C}$  was not much effected by the outliers. The diagonal elements are near 1 and the off-diagonal elements are near 0. Figure 4.13 shows that the sample covariance matrix  $\mathbf{S}$  was greatly effected by the outliers. Several sample covariances are less than  $-20$ , and several sample variances are over 40.



**Fig. 4.13** Elements of the classical covariance matrix  $S$  for outlier data

*R* code to used to produce Figures 4.12 and 4.13 is shown below.

```
#n = 50, p = 100
x<-matrix(rnorm(5000),nrow=50,ncol=100)
out<-medout(x) #no outliers, try ddplot5(x)
out <- covmb2(x,msteps=0)
z<-out$cov
plot(diag(z)) #plot the diagonal elements of C
plot(out$center) #plot the elements of T
vecz <- vecw(z)$vecz
plot(vecz)

out<-covmb2(x,m=45)
plot(out$center)
plot(diag(out$cov))

#outliers
x[1:10,] <- 10*x[1:10,]
x[1:10,1] <- x[1:10]+10
medout(x) #The 10 outliers are easily detected in
#the plot of the distances from the MED(X).
ddplot5(x) #two widely separated clusters of data
tem <- getB(x,msteps=0)
tem$indx #all 40 clean cases were used
dim(tem$B) #40 by 100
out<-covmb2(x,msteps=0)
```

```

z<-out$cov
plot(diag(z))
plot(out$center)
vecz <- vecw(z)$vecz
plot(vecz) #plot the elements of C
#Figure 4.12

#examine the sample covariance matrix and mean
plot(diag(var(x)))
plot(apply(x,2,mean)) #plot elements of xbar
zc <- var(x)
vecz <- vecw(zc)$vecz
plot(vecz) #plot the elements of S
#Figure 4.13

out<-medout(x) #10 outliers
out<-covmb2(x,m=40)
plot(out$center)
plot(diag(out$cov))

```

The `covmb2` estimator can also be used for  $n > p$ . The *mpack* function `mldsim6` suggests that for 40% outliers, the outliers need to be further away from the bulk of the data (`covmb2(k=5)` needs a larger value of  $pm$ ) than for the other six estimators if  $n \geq 20p$ . With outlier types like those in Tables 4.4 and 4.5, `covmb2(k=5)` was often near best. Try the following commands. The other estimators need  $n > 2p$ , and as  $n$  gets close to  $2p$ , `covmb2` may outperform the other estimators.

```

#near point mass on major axis
mldsim6(n=100,p=10,outliers=1,gam=0.25,pm=25)
mldsim6(n=100,p=10,outliers=1,gam=0.4,pm=25) #bad
mldsim6(n=100,p=40,outliers=1,gam=0.1,pm=100)
mldsim6(n=200,p=60,outliers=1,gam=0.1,pm=100)
#mean shift outliers
mldsim6(n=100,p=40,outliers=3,gam=0.1,pm=10)
mldsim6(n=100,p=40,outliers=3,gam=0.25,pm=20)
mldsim6(n=200,p=60,outliers=3,gam=0.1,pm=10)
#concentration steps can help
mldsim6(n=100,p=10,outliers=3,gam=0.4,pm=10,osteps=0)
mldsim6(n=100,p=10,outliers=3,gam=0.4,pm=10,osteps=9)

```

The following estimator can also be used if  $p > n$ , but improvements are needed, and suggested in the following paragraphs. For other estimators that can be used when  $p > n$ , see Boudt et al. (2017).

**Definition 4.13.** The Sign Covariance Matrix

$$\hat{\Sigma}_S = \frac{1}{n} \sum_{i=1}^n \frac{(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_n)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_n)^T}{\|\mathbf{x}_i - \hat{\boldsymbol{\mu}}_n\|^2} = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i^T.$$

This estimator is similar to the classical covariance estimator computed from  $\mathbf{z}_i = \frac{\mathbf{x}_i - \hat{\boldsymbol{\mu}}_n}{\|\mathbf{x}_i - \hat{\boldsymbol{\mu}}_n\|}$ , assuming  $\bar{\mathbf{z}}_i \approx \mathbf{0}$ . Note that  $\|\mathbf{z}_i\| = 1$ , so the  $\mathbf{z}_i$  lie on the unit hypersphere centered at  $\hat{\boldsymbol{\mu}}_n$ . Here  $\hat{\boldsymbol{\mu}}_n$  is the  $L_1$ -median or spatial median, defined as

$$\hat{\boldsymbol{\mu}}_n = \operatorname{argmin}_{\boldsymbol{\mu}} \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i - \boldsymbol{\mu}\|,$$

and  $\hat{\boldsymbol{\mu}}_n$  is a fairly practical high breakdown estimator of multivariate location. An argument similar to the proof of Lemma 4.3 can be used to show that the maximum eigenvalue of  $\hat{\Sigma}_S$  is bounded.

Draw a circle and then ellipsoidal data centered at the center of the circle. Project the data on the circle. Unless the data is spherical, the projection severely distorts the shape of the data. Hence  $\hat{\Sigma}_S$  is not a consistent estimator of  $c\boldsymbol{\Sigma}$  for nonspherical elliptically contoured data. Suppose  $p = 2$  and the highest density region is an ellipse with the  $x$ -axis as the major axis and the  $y$ -axis as the minor axis. Assume the data projected on the  $x$ -axis has standard deviation 3, and the data projected on the  $y$ -axis has standard deviation 0.3. Then we expect that  $\hat{\Sigma}_S$  will underestimate the variability about the major axis and overestimate the variability about the minor axis.

Suppose  $p$  is fixed and  $n \rightarrow \infty$ . Croux et al. (2010) showed that the Sign Covariance Matrix is a high breakdown estimator, and the minimum eigenvalue can be driven to zero if more than half of the cases are outliers. They also claim, that under regularity conditions, that for clean data, the Sign Covariance Matrix consistently estimates the orientation of the dispersion matrix: for a class of elliptically contoured distributions, the eigenvectors  $\hat{\mathbf{e}}_i$  estimate the population eigenvectors  $\mathbf{e}_i$ .

**Remark 4.6.** Suppose the  $\hat{\mathbf{e}}_i$  estimate the dispersion matrix orientation, but the eigenvalues are severely biased. Let  $L_i = \operatorname{MAD}(\hat{\mathbf{e}}_i^T \mathbf{x}_1, \dots, \hat{\mathbf{e}}_i^T \mathbf{x}_n)$  for  $i = 1, \dots, p$ . Note that  $\hat{\mathbf{e}}_i^T \mathbf{x}_1, \dots, \hat{\mathbf{e}}_i^T \mathbf{x}_n$  are the data projected onto the line in the direction of the  $i$ th orthonormal eigenvector  $\hat{\mathbf{e}}_i$ . Let  $\hat{\lambda}_1 = L_{(p)} \geq \hat{\lambda}_2 = L_{(p-1)} \geq \dots \geq \hat{\lambda}_{p-1} = L_{(2)} \geq \hat{\lambda}_p = L_{(1)}$ . If necessary, relabel the  $\hat{\mathbf{e}}_i$  so that  $\hat{\lambda}_i$  corresponds to eigenvector  $\hat{\mathbf{e}}_i$ . Then let  $\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^p \hat{\lambda}_i \hat{\mathbf{e}}_i \hat{\mathbf{e}}_i^T$ . (Compare Theorem 2.1.) Then on a class of elliptically contoured distributions,  $\hat{\boldsymbol{\Sigma}}$  may be a better estimator of  $c\boldsymbol{\Sigma}$  for some  $c > 0$ .

Outliers can still severely bias the estimator since they have the same weight as the clean data. For example, draw a circle centered at the origin. Suppose the data  $\mathbf{x}_i$  is tightly clustered about the horizontal axis. Then the projected data  $\mathbf{z}_i$  tightly clusters about the points  $(-1, 0)$  and  $(1, 0)$ . Placing outliers with  $\mathbf{z}_i$  at  $(0, 1)$  or  $(0, -1)$  will inflate the minor axis.

To examine breakdown, suppose that more than half of the  $\mathbf{x}_i = \mathbf{x}_o$ , so there is a point mass at  $\mathbf{x}_o$ . Because of the exact fit property, the spatial median is equal to  $\mathbf{x}_o$ . Draw a cloud for the clean data and put  $\mathbf{x}_o$  far to the right of the cloud. Draw a unit circle around  $\mathbf{x}_o$  and a left opening cone with vertex at  $\mathbf{x}_o$  that just contains the data cloud. Then the clean data gets projected on an arc where the cone intersects the left of the circle. As  $\mathbf{x}_o$  is moved further to the right, the arc length goes to zero, so the projected data form a near point mass. Hence the Sign Covariance Matrix is roughly the covariance matrix of a point mass (it is not clear what  $\mathbf{z}_i$  is if  $\mathbf{x}_i = \hat{\boldsymbol{\mu}}_n$ ) and a near point mass, which will have smallest eigenvalue  $\rightarrow 0$  as  $\mathbf{x}_o$  gets further away from the clean data cloud.

To examine the effect of outliers on the eigenvectors, assume the proportion of outliers  $\mathbf{x}_o$  is 40%. Then the spatial median gets dragged to the right of the data cloud. Draw a unit circle around the spherical median and project the data onto the unit circle. Most of the data is projected onto the left half of the circle, so the eigenvectors  $\hat{\mathbf{e}}_i$  no longer estimate the orientation of the dispersion matrix.

A simple solution is to compute the spatial median on the `covmb2` set  $B$ , perhaps using the spatial median instead of the coordinatewise median to form the  $D_i$ , but it seems that applying the classical estimator on the data in set  $B$  makes more sense.

The median ball estimator and `covmb2` put a hypersphere, centered at the coordinatewise median, about the data and gives zero weight to data outside the hypersphere. The Sign Covariance Matrix projects the data onto a unit hypersphere centered at the spherical median. We conjecture that the median ball and `covmb2` estimators also estimate the orientation of the dispersion matrix for a class of elliptically contoured distributions. The reweighted median ball estimator with five concentration steps seems to estimate the  $c\boldsymbol{\Sigma}$  with very small bias for many elliptically contoured distributions when  $n \gg p$ . When  $p > n$  and `covmb2` is used, using Remark 4.6 to find the  $L_i$  and  $\hat{\lambda}_i$  may be useful.

## 4.8 Summary

1) Given a table of data  $\mathbf{W}$  for variables  $X_1, \dots, X_p$ , be able to find the **coordinatewise median**  $\text{MED}(\mathbf{W})$  and the **sample mean**  $\bar{\mathbf{x}}$ . If  $\mathbf{x} = (X_1, X_2, \dots, X_p)^T$  where  $X_j$  corresponds to the  $j$ th column of  $\mathbf{W}$ , then  $\text{MED}(\mathbf{W}) = (\text{MED}_{X_1}(n), \dots, \text{MED}_{X_p}(n))^T$  where  $\text{MED}_{X_j}(n) = \text{MED}(X_{j,1}, \dots,$

$X_{j,n}$ ) is the sample median of the data in the  $j$ th column. Similarly,  $\bar{\mathbf{x}} = (\bar{X}_1, \dots, \bar{X}_p)^T$  where  $\bar{X}_j$  is the sample mean of the data in the  $j$ th column.

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

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

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

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

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

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

$$\mathbf{C}_{FCH} = \frac{\text{MED}(D_i^2(T_A, \mathbf{C}_A))}{\chi_{p,0.5}^2} \mathbf{C}_A \quad (4.12)$$

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



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

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

10) Brand-name robust MLD estimators from the dominant robust statistics paradigm take too long to compute: F-brand-name estimators that are not backed by breakdown or large sample theory are actually used. FMCD, F-MVE, F-S, F-MM, F- $\tau$ , F-constrained-M, and F-Stahel–Donoho are especially common.

## 4.9 Complements

Most of this chapter focused on robust estimators where  $n \geq 10p$ . Dispersion estimators for  $n < 10p$  are discussed in Section 4.7, Pourahmadi (2013), and Yao et al. (2015). Tiao and Tsay (1983) gave an interesting bound on the determinant.

For concentration algorithms, note that  $(T_{t,j}, \mathbf{C}_{t,j}) = (\bar{\mathbf{x}}_{t,j}, \mathbf{S}_{t,j})$  is the classical estimator applied to the “half set” of cases satisfying  $\{\mathbf{x}_i : D_i^2(\bar{\mathbf{x}}_{t-1,j}, \mathbf{S}_{t-1,j}) \leq D_{(c_n)}^2(\bar{\mathbf{x}}_{t-1,j}, \mathbf{S}_{t-1,j})\}$  for  $t \geq 0$ . Hence  $(T_{t,j}, \mathbf{C}_{t,j})$  is estimating  $(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ , the population mean and covariance matrix of the truncated distribution covering half of the mass corresponding to  $\{\mathbf{x} : (\mathbf{x} - \boldsymbol{\mu}_{t-1})^T \boldsymbol{\Sigma}_{t-1}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{t-1}) \leq D_{0.5}^2(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})\}$  where  $D_{0.5}^2(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})$  is the population median of the population squared distances  $D^2(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})$ . Here  $(\boldsymbol{\mu}_{-1}, \boldsymbol{\Sigma}_{-1})$  is the population analog of  $(T_{-1,j}, \mathbf{C}_{-1,j})$ .

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

**Conjecture 4.2.** The DGK estimator is a  $\sqrt{n}$  consistent estimator of  $(\boldsymbol{\mu}_{k,D}, \boldsymbol{\Sigma}_{k,D})$  under mild conditions.

The median ball (MB) estimator  $(T_{k,M}, \mathbf{C}_{k,M})$  uses  $(T_{-1,M}, \mathbf{C}_{-1,M}) = (\text{MED}(\mathbf{W}), \mathbf{I}_p)$  as the only start where  $\text{MED}(\mathbf{W})$  is the coordinatewise median. Hence  $(T_{0,M}, \mathbf{C}_{0,M})$  is the classical estimator applied to the “half

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

**Conjecture 4.3.** The MB estimator is a high breakdown  $\sqrt{n}$  consistent estimator of  $(\boldsymbol{\mu}_{k,M}, \boldsymbol{\Sigma}_{k,M})$  under mild conditions. For elliptically contoured distributions,  $\boldsymbol{\mu}_{k,M} = \boldsymbol{\mu}$ .

The `covmb2` estimator is also a natural estimator of a population mean and covariance matrix  $(\boldsymbol{\mu}_{j,B}, \boldsymbol{\Sigma}_{j,B})$  corresponding to a truncated distribution, and we conjecture that `covmb2` is  $\sqrt{n}$  consistent estimator of  $(\boldsymbol{\mu}_{j,B}, \boldsymbol{\Sigma}_{j,B})$  under mild conditions.

Arcones (1995) and Kim (2000) showed that  $\bar{\mathbf{x}}_{0,M}$  is a HB  $\sqrt{n}$  consistent estimator of  $\boldsymbol{\mu}$ . Olive (2004a) showed that  $(\bar{\mathbf{x}}_{0,M}, \mathbf{S}_{0,M}) = (T_{0,M}, \mathbf{C}_{0,M})$  is a high breakdown estimator. If the data distribution is EC but not spherical about  $\boldsymbol{\mu}$ , then for  $k \geq 0$ ,  $\mathbf{S}_{k,M} = \mathbf{C}_{MB}$  underestimates the major axis and overestimates 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.

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  $\text{MED}(\mathbf{W})$  is the sample median of the  $j$ th predictor. Hence with  $m - 1$  outliers,  $\text{MED}(\mathbf{W})$  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  $\text{MED}(\mathbf{W})$  can be moved to the boundary of the much smaller “coordinatewise IQR box” corresponding the 25th and 75th percentiles of the clean data. Then the edge lengths are approximately equal to the interquartile ranges of the clean cases.

Note that  $D_i(\text{MED}(\mathbf{W}), \mathbf{I}_p) = \|\mathbf{x}_i - \text{MED}(\mathbf{W})\|$  is the Euclidean distance of  $\mathbf{x}_i$  from  $\text{MED}(\mathbf{W})$ . 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

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

One of the most effective methods for detecting outliers for large data sets or if  $p > n$  is to use  $D_i(\text{MED}(\mathbf{W}), \mathbf{I}_p)$ . See the *mpack* function `medout`. Section 4.7 suggests more useful techniques.

The MB estimator has outlier resistance similar to  $(\text{MED}(\mathbf{W}), \mathbf{I}_p)$  for distant outliers but, as shown in Example 4.1, can be much more effective for detecting certain types of outliers that cannot 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  $\{\mathbf{x}_i : D_i^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leq d^2\}$ .

If the DGK estimator is used by itself, we recommend  $k = 10$  in the concentration algorithm. We use  $k = 5$  when the DGK and MB estimators are used as attractors for the FCH, CMVE, and MBA estimators. The scaling (4.10) makes  $\mathbf{C}_{FCH}$  a better estimate of  $\boldsymbol{\Sigma}$  if the data is multivariate normal MVN.

Concentration for the MB estimator begins with the “half set” of data closest to the coordinatewise median in Euclidean distance, resulting in the estimator  $(T_{0,M}, \mathbf{C}_{0,M})$  that uses 50% trimming.  $(T_{0,M}, \mathbf{C}_{0,M})$  is a high breakdown estimator by Corollary 4.7. Since only cases  $\mathbf{x}_i$  such that  $\|\mathbf{x}_i - \text{MED}(\mathbf{W})\| \leq \text{MED}(\|\mathbf{x}_i - \text{MED}(\mathbf{W})\|)$  are used, the largest eigenvalue of  $\mathbf{C}_{0,50}$  is bounded if fewer than half of the cases are outliers by Lemma 4.3.

The geometric behavior of  $(T_{0,M}, \mathbf{C}_{0,M})$  is simple. If the data  $\mathbf{x}_i$  are MVN (or EC with continuous decreasing  $g$ ) then the highest density regions of the data are hyperellipsoids. The set of  $\mathbf{x}$  closest to the coordinatewise median in Euclidean distance is a hypersphere. For EC data, the highest density hyperellipsoid will have approximately the same center as the hypersphere, and the hypersphere will be drawn toward the longest axis of the hyperellipsoid. Hence too much data will be trimmed in that direction. For example, if the data are MVN with  $\boldsymbol{\Sigma} = \text{diag}(1, 2, \dots, p)$  then  $\mathbf{C}_{0,M}$  will underestimate the largest variance and overestimate the smallest variance. Taking  $k$  concentration steps can greatly reduce but not eliminate the bias of the MB estimator  $\mathbf{C}_{k,M}$  if the data is EC, and the determinant  $|\mathbf{C}_{k,M}| < |\mathbf{C}_{0,M}|$  unless the attractor is equal  $(T_{0,M}, \mathbf{C}_{0,M})$  by Proposition 4.4. The MB estimator  $(T_{k,M}, \mathbf{C}_{k,M})$  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) suggested an estimator similar to the MB estimator, also see Croux and Van Aelst (2002). Another estimator similar to MB was suggested by Wilk et al. (1962). See Gnanadesikan (1977, p. 134).

Recall that the *population squared Mahalanobis distance*

$$U \equiv D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}). \quad (4.13)$$

For elliptically contoured distributions,  $U$  has pdf given by (3.10), and if  $g$  is continuous and decreasing, then the 50% highest density region has the form of the hyperellipsoid

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

where  $u_{0.5}$  is the median of the distribution of  $U$ . For example, if the  $\mathbf{x}$  are MVN, then  $U$  has the  $\chi_p^2$  distribution. Concentration estimators attempt to estimate the population mean and covariance matrix of the mass in this 50% highest density region. So it should not be surprising that good concentration attractors estimate the same quantity  $(\mu, a_{MCD}\Sigma)$ . See Theorem 4.9.

In regression, if the start is a consistent estimator for  $\beta$ , then so is the attractor. Hence all attractors are estimating the *same* parameter  $\beta$ . Theorem 4.9 showed that MLD concentration attractors with  $k \geq 0$  are estimating the *same* parameter  $(\mu, a_{MCD}\Sigma)$  even if the affine equivariant starts are estimating  $(\mu, s_i\Sigma)$  where the  $s_i > 0$  can differ for  $i = 1, \dots, K$ .

Olive (2002) showed the following result. Assume  $(T_i, \mathbf{C}_i)$  are consistent estimators for  $(\mu, a_i\Sigma)$  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  $D_i(T_1, \mathbf{C}_1) \leq \text{MED}(D_i(T_1, \mathbf{C}_1))$ . Let  $r_n$  be the correlation between  $D_{i,1}$  and  $D_{i,2}$  for the cases in  $R$ . Then  $r_n \rightarrow 1$  in probability as  $n \rightarrow \infty$ .

The theory for concentration algorithms is due to Hawkins and Olive (2002) and Olive and Hawkins (2010). The MBA estimator is due to Olive (2004a). The computational and theoretical simplicity of the FCH estimator makes it interesting. 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.

The recommended estimators are the RFCH and RMVN estimators. These two estimators are about an order of magnitude faster than most alternative robust estimators, but take slightly longer to compute than the FCH estimator, and may have slightly less resistance to outliers. RFCH and RMVN appear to have much less variability than FCH when there are no outliers. These three estimators are in Zhang et al. (2012). Also see Zhang and Olive (2009).

In addition to concentration and randomly selecting elemental sets, three other algorithm techniques are important. He and Wang (1996) suggested 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  $\mathbf{C} = (c_{i,j})$  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

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

where  $\text{VAR}(X) = \sigma^2(X)$  suggests that a robust estimator of dispersion can be created by replacing the sample standard deviation  $\hat{\sigma}$  by a robust estimator of scale. (This idea seems best if the outlying cases have outliers that can be detected if projected on two of the  $p$  coordinate axes. It is possible to have outliers that do not appear in one- or two-dimensional projections. Hence the method may not perform well compared to methods that use Mahalanobis or Euclidean distances like FCH and `covmb2`.) Maronna and Zamar (2002) modified this idea to create a fairly fast (possibly high breakdown consistent) OGK estimator of multivariate location and dispersion. Also see Alqallaf et al. (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).

Billor et al. (2000) proposed a BACON algorithm that uses  $m_0 = 4p$  or  $m_0 = 5p$  cases, computes the sample mean and covariance matrix of these cases, finds the  $m_1$  cases with Mahalanobis distances less than some cutoff, then iterates until the subset of cases no longer changes. Version V1 uses the  $m_0$  cases with the smallest classical Mahalanobis distances while version V2 uses the  $m_0$  cases closest to the coordinatewise median.

There certainly exist types of outlier configurations where the FMCD estimator outperforms the robust RFCH estimator. The RFCH estimator is vulnerable to certain types of outliers that lie inside the hypersphere based on the median Euclidean distance from the coordinatewise median. The FMCD estimator should be modified so that it is backed by theory. 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, RFCH, 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 4.7. The Det-MCD estimator of Hubert et al. (2002) is very similar, uses 6 starts, but is not yet backed by theory.

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

Rocke and Woodruff (1996, p. 1050) claimed 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 et al. (1993) for an exact algorithm for the MVE. Other papers on

robust algorithms include Hawkins (1993, 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 described many methods.

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

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

Huber and Ronchetti (2009, pp. 214, 233) noted that theory for M estimators of multivariate location and dispersion is “not entirely satisfactory with regard to joint estimation of”  $(\mu, a\Sigma)$  and that “so far we have neither a really fast, nor a demonstrably convergent, procedure for calculating simultaneous  $M$ -estimates of location and scatter.”

If an exact algorithm exists but an approximate algorithm is also used, the two estimators should be distinguished in some manner. For example,  $(T_{MCD}, \mathbf{C}_{MCD})$  could denote the estimator from the exact algorithm while  $(T_{FMCD}, \mathbf{C}_{FMCD})$  could denote the estimator from the approximate algorithm. In the literature, this distinction is too seldomly made, but there are a few outliers. Cook and Hawkins (1990, p. 640) pointed out that the FMVE is not the minimum volume ellipsoid (MVE) estimator.

### Where the Dominant Robust Statistics Paradigm Goes Wrong

i) Estimators from this paradigm that have been shown to be both high breakdown and consistent take too long to compute.

Let the  $i$ th case  $\mathbf{x}_i$  be a  $p \times 1$  random vector, and suppose the  $n$  cases are collected in an  $n \times p$  matrix  $\mathbf{W}$  with rows  $\mathbf{x}_1^T, \dots, \mathbf{x}_n^T$ . The fastest estimators of multivariate location and dispersion that have been shown to be both consistent and high breakdown are the minimum covariance determinant (MCD) estimator with  $O(n^v)$  complexity where  $v = 1 + p(p+3)/2$  and possibly an all elemental subset estimator of He and Wang (1997). See Bernholt and Fischer (2004). The minimum volume ellipsoid complexity is far higher, and for  $p > 2$  **there may be no known method for computing** S,  $\tau$ , projection based, and constrained M estimators. For some depth estimators, like the Stahel–Donoho estimator, the exact algorithm of Liu and Zuo (2014) appears to take too long if  $p \geq 6$  and  $n \geq 100$ , and simulations may need  $p \leq 3$ .

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

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

ii) No practical useful “high breakdown” estimator of multivariate location and dispersion (MLD) from this paradigm has been shown to be both consistent and high breakdown: to my knowledge, **if the complexity of the estimator is less than  $O(n^4)$  for general  $p$ , and if the estimator has been claimed in the published literature to be both high breakdown and consistent, then the MLD estimator has not been shown to be both high breakdown and consistent.** Also Hawkins and Olive (2002) showed that elemental concentration estimators using  $K$  starts are zero breakdown estimators, and these estimators are inconsistent if they use  $k$  concentration steps where  $k$  is fixed.

The **main competitors** for the Olive and Hawkins (2010) multivariate location and dispersion FCH, RFCH, and RMVN estimators are the Maronna and Zamar (2002) *OGK estimator*, the Hubert et al. (2012) *Det-MCD estimator* which have not been proven to be consistent or positive breakdown, and the *Sign Covariance Matrix* shown to be high breakdown by Croux et al. (2010). Also see Taskinen et al. (2012). Croux et al. (2010) showed that the practical Sign Covariance Matrix and  $k$ -step Spatial Sign Covariance Matrix are high breakdown. They claimed that under regularity conditions, these two estimators consistently estimate the orientation of the scatter matrix. Sections 4.5 and 4.7 suggest that a tight cluster of outliers severely bias these competing estimators, unless the outlier proportion is small.

Papers with titles like Rousseeuw and Van Driessen (1999) “A Fast Algorithm for the Minimum Covariance Determinant Estimator” and Hubert et al. (2008) “High Breakdown Multivariate Methods” where the zero breakdown estimators have not been shown to be consistent are common, and very misleading to researchers who are not experts in robust statistics.

iii) Many papers give theory for an impractical estimator such as MCD, then replace the estimator by a zero breakdown practical estimator such as FMCD.

It may be reasonable to call an estimator a brand-name high breakdown estimator (e.g., S estimator or MCD estimator) if a) the estimator is high breakdown and b) the estimator has the same asymptotic distribution as the brand-name estimator.

Since the brand-name estimators are impractical to compute, practical algorithm estimators that use a fixed number of trial fits are used. If a) and b) have not been proven for the practical estimator, call the practical estimator an F-brand-name estimator (e.g., FS or FMCD), where F denotes



that the criterion of the brand-name estimator was used to select a trial fit to be used in the F-brand-name estimator from a fixed number of trial fits.

The F-brand-name estimators can have a wide variety of theory. Suppose the final estimator is one of the trial fits. a) If  $K$  elemental sets are used as the trial fits then the final estimator is inconsistent and zero breakdown by Theorem P.1 in the preface. b) If there is only one trial fit, computing the criterion does not change the trial fit. Then the final estimator is the trial fit. So if the classical estimator is the trial fit, then the classical estimator is the final estimator. c) If  $K$  trial fits are used and each trial fit is a  $\sqrt{n}$  consistent estimator of  $(\mu, c\Sigma)$ , then the final estimator is a  $\sqrt{n}$  consistent estimator of  $(\mu, c\Sigma)$ .

A necessary condition for a practical estimator to be the global minimizer of a brand-name criterion is that the practical estimator has a criterion value at least as small as any other estimator used. So if the FMCD estimator is the MCD estimator,  $\det(\mathbf{C}_{FMCD}) \leq \det(\mathbf{C}_A)$  for any other estimator  $\mathbf{C}_A$  that is the sample covariance matrix applied to the same number of cases  $c$  as the FMCD estimator. Hubert et al. (2012) claimed the Fast-MCD is the MCD estimator, but sometimes Fast-MCD has the smallest determinant and sometimes Det-MCD has the smallest determinant. This result proves that neither Fast-MCD nor Det-MCD is computing MCD. Similarly, permuting the data should not change the criterion if the global minimizer is being computed. See Problem 4.9.

To get theory for Det-MCD, we need breakdown results when the concentration steps are iterated to convergence, and we need a result like Theorem 4.9 when the initial estimator is consistent but not necessarily affine equivariant, and the initial estimator is iterated to convergence.

iv) Papers on breakdown and maximal bias are not useful.

Both these properties are weaker than asymptotic unbiasedness. Also the properties are derived for estimators that take far too long to compute.

Breakdown is a very weak property: having  $\|T\|$  bounded and eigenvalues of  $\mathbf{C}$  bounded away from 0 and  $\infty$  does not mean that the estimator is good. All too often claims are made that “high breakdown estimators make outliers have large distances.”

Sometimes the literature gives a claim similar to “the fact that FMCD is not the MCD estimator is unimportant since the algorithm that uses all elemental sets has the same high breakdown value as MCD.” FMCD is not the MCD estimator and FMCD is not the estimator that uses all elemental sets. FMCD only uses a fixed number of elemental sets, hence FMCD is zero breakdown.

v) Too much emphasis is given on the property of affine equivariance since typically this is the only property that can be shown for a practical estimator of MLD.

Huber and Ronchetti (2009, pp. 200, 283) noted that “one ought to be aware that affine equivariance is a requirement deriving from mathematical aesthetics; it is hardly ever dictated by the scientific content of the underly-



ing problem,” and the lack of affine equivariance “may be less of a disadvantage than it first seems, since in statistics problems possessing genuine affine equivariance are quite rare.” Also see the warning at the end of Section 4.1.

Being a  $\sqrt{n}$  consistent estimator of  $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$  is an important property, and the FCH estimator is asymptotically equivalent to the scaled DGK estimator, which is affine equivariant.

The major algorithm producers are now using intelligently chosen starts or trial fits, such as the classical estimator, in their F-brand-name estimators. These estimators are typically not affine equivariant. See Hawkins and Olive (2002), Olive (2004a), Olive and Hawkins (2010), Hubert et al. (2012), and Maronna and Yohai (2015).

vi) The literature implies that the breakdown value is a measure of the global reliability of the estimator and is a lower bound on the amount of contamination needed to destroy an estimator.

These interpretations are not correct since the complement of complete and total failure is *not* global reliability. The breakdown value  $d_n/n$  is actually an upper bound on the amount of contamination that the estimator can tolerate since the estimator can be made arbitrarily bad with  $d_n$  maliciously placed cases. In particular, the breakdown value of an estimator tells nothing about more important properties such as consistency or asymptotic normality.

## 4.10 Problems

**PROBLEMS WITH AN ASTERISK \* ARE ESPECIALLY USEFUL.**

### R Problems

Use the command `source("G:/mpack.txt")` to download the functions and the command `source("G:/mrobddata.txt")` to download the data. See Preface or Section 15.2. Typing the name of the `mpack` function, e.g., `covmba`, will display the code for the function. Use the `args` command, e.g., `args(covmba)`, to display the needed arguments for the function. For some of the following problems, the *R* commands can be copied and pasted from (<http://lagrange.math.siu.edu/Olive/mrsashw.txt>) into *R*.

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

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

**4.2.** Download the `rmaha` function that creates the robust Mahalanobis distances using `cov.mcd` (FMCD). Obtain `outx2` as in Problem 4.1b). Enter the *R* command `library(MASS)`. Enter the command `rmaha(outx2)` and check whether observations 1–40 look like outliers.

**4.3.** 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.

**4.4\*.** a) Assuming that you have done the two source commands above Problem 4.1 (and the *R* command `library(MASS)`), type the command `ddcomp(buxx)`. This will make 4 DD plots based on the DGK, FCH, FMCD, and median ball estimators. The DGK and median ball estimators are the two attractors used by the FCH estimator. With the leftmost mouse button, move the cursor to an outlier and click. This data is the Buxton (1920) data and cases with numbers 61, 62, 63, 64, and 65 were the outliers with head lengths near 5 feet. After identifying at least three outliers in each plot, hold the rightmost mouse button down (and in *R* click on *Stop*) to advance to the next plot. When done, hold down the *Ctrl* and *c* keys to make a copy of the plot. Then paste the plot in *Word*.

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

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

**4.5\*.** (Perform the `source("G:/mpack.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 robust estimators cannot always detect them. Type the command `concmv()`. Hold the rightmost mouse button down (and in *R* click on *Stop*) to see the DD plot after one concentration step. The start uses the coordinatewise median and  $\text{diag}([MAD(X_i)]^2)$ . Repeat four 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*.

**4.6\*.** (Perform the `source("G:/mpack.txt")` command if you have not already done so.) The `ddmv` function illustrates concentration with the DD plot. The outliers are highlighted. The first graph is the DD plot after one concentration step. Hold the rightmost mouse button down (and in *R* click on *Stop*) to see the DD plot after two concentration steps. Repeat four 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  $0 < \text{gam} < 0.5$  as large as possible.

**4.7.** (Perform the `source("G:/mpack.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  $\text{MED}(\mathbf{x})$ . For example, use  $M \in \{98, 95, 90, 80, 70, 60, 50\}$ . Obtain the program `cmba2` from `mpack.txt` and try the MBA estimator on the data sets in Problem 4.4. You need to right click *Stop* on the plot 7 times.

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

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

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

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

a) Paste the output into *Word*.

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

Robust Multivariate Analysis

J. Olive, D.

2017, XVI, 501 p. 76 illus., 6 illus. in color., Hardcover

ISBN: 978-3-319-68251-8