Chapter 4

MLD Estimators

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 pth 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 $(T, \boldsymbol{C})$ 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 σ_{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 $X_1, ..., X_p$.

Rule of thumb 4.1. For the classical estimators of multivariate location and dispersion, $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ or $(\overline{\boldsymbol{z}}, \boldsymbol{R})$, want $n > 10p$. Want $n > 20p$ 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 $\boldsymbol{W}$. Let

$\boldsymbol{B} = \mathbf{1}\boldsymbol{b}^T$ where $\mathbf{1}$ is an $n \times 1$ vector of ones and $\boldsymbol{b}$ is a $p \times 1$ constant vector. Hence the ith row of $\boldsymbol{B}$ is $\boldsymbol{b}_i^T \equiv \boldsymbol{b}^T$ for $i = 1, ..., n$. For such a matrix $\boldsymbol{B}$, consider the affine transformation $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{A} + \boldsymbol{B}$ where $\boldsymbol{A}$ is any nonsingular $p \times p$ matrix.

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

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

and

$$\boldsymbol{C}(\boldsymbol{Z}) = \boldsymbol{C}(\boldsymbol{W}\boldsymbol{A} + \boldsymbol{B}) = \boldsymbol{A}^T \boldsymbol{C}(\boldsymbol{W})\boldsymbol{A}. \tag{4.2}$$

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

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

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

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

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

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

Warning: Estimators that use randomly chosen elemental sets or projections are not affine equivariant since these estimators change every time they are computed. Such estimators can sometimes be made affine equivariant by using the same fixed random number seed each time the estimator is used. Then the affine equivariance of the estimator depends on the random number seed, and such estimators are not as attractive as affine equivariant estimators that do not depend on a fixed random number seed.

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 $\boldsymbol{W}$ denote the $n \times p$ data matrix with ith row $\boldsymbol{x}_i^T$ corresponding to the ith case. Let $\boldsymbol{w}_1, ...\boldsymbol{w}_n$ be the contaminated data after d_n of the $\boldsymbol{x}_i$ have been replaced by arbitrarily bad contaminated cases. Let $\boldsymbol{W}_d^n$ denote the $n \times p$ data matrix with ith row $\boldsymbol{w}_i^T$. Then the contamination fraction is $\gamma_n = d_n/n$. Let $(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W}))$ denote an estimator of multivariate location and dispersion where the $p \times 1$ vector $T(\boldsymbol{W})$ is an estimator of location and the $p \times p$ symmetric positive semidefinite matrix $\boldsymbol{C}(\boldsymbol{W})$ is an estimator of dispersion. Recall from Theorem 1.1 that if $\boldsymbol{C}(\boldsymbol{W}_d^n) > 0$, then $\max_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^T \boldsymbol{C}(\boldsymbol{W}_d^n)\boldsymbol{a} = \lambda_1$ and $\min_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^T \boldsymbol{C}(\boldsymbol{W}_d^n)\boldsymbol{a} = \lambda_p$. A high breakdown dispersion estimator $\boldsymbol{C}$ is positive definite if the amount of contamination is less than the breakdown value. Since $\boldsymbol{a}^T \boldsymbol{C}\boldsymbol{a} = \sum_{i=1}^p \sum_{j=1}^p c_{ij} a_i a_j$, the largest eigenvalue λ_1 is bounded as $\boldsymbol{W}_d^n$ varies iff $\boldsymbol{C}(\boldsymbol{W}_d^n)$ is bounded as $\boldsymbol{W}_d^n$ varies.

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

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

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

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

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

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

The following result shows that a multivariate location estimator T basically "breaks down" if the d outliers can make the median Euclidean distance $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|)$ arbitrarily large where $\boldsymbol{w}_i^T$ is the ith row of $\boldsymbol{W}_d^n$. Thus a multivariate location estimator T will not break down if T can not be driven out of some ball of (possibly huge) radius r about the origin.

Proposition 4.2. 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 is the smallest number of arbitrarily bad cases that can make the median Euclidean distance $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|)$ arbitrarily large.

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

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

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

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

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

maximum eigenvalue λ_1 of $\boldsymbol{S}_B$ is bounded.

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

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

Hence the maximum eigenvalue λ_1 is bounded. □

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

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq D^2_{(c_n)}\} \tag{4.4}$$

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

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

has volume equal to

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

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

4.3 The Concentration Algorithm

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 $\boldsymbol{C}_{MCD}$ denote the sample mean and sample covariance matrix of the c_n cases in J_o. Then the *minimum covariance determinant* MCD(c_n) estimator is $(T_{MCD}(\boldsymbol{W}), \boldsymbol{C}_{MCD}(\boldsymbol{W}))$.

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

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

where a_{MCD} is some positive constant when the data $\boldsymbol{x}_i$ are elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$, and T_{MCD} has a Gaussian limit. See Butler, Davies, and Jhun (1993) and Cator and Lopuhaä (2009, 2010).

Computing robust covariance estimators can be very expensive. For example, to compute the exact MCD(c_n) estimator (T_{MCD}, C_{MCD}), we need to consider the $C(n, c_n)$ subsets of size c_n. Woodruff and Rocke (1994, p. 893) note that if 1 billion subsets of size 101 could be evaluated per second, it would require 10^{33} millenia to search through all $C(200, 101)$ subsets if the sample size $n = 200$.

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

algorithm estimator is a special case where $k = -1$ so that the attractor is the start: $(\overline{\boldsymbol{x}}_{k,j}, \boldsymbol{S}_{k,j}) = (\overline{\boldsymbol{x}}_{-1,j}, \boldsymbol{S}_{-1,j})$.

This concentration algorithm is a simplified version of the algorithms given by Rousseeuw and Van Driessen (1999) and Hawkins and Olive (1999). Using $k = 10$ concentration steps often works well.

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

Starts that use a consistent initial estimator could be used. K_n is the number 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, eg $K = 500$, so K does not depend on n. A crucial observation is that the theory of the algorithm estimator depends on the theory of the attractors, not on the estimator corresponding to the criterion.

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

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

This theory shows that the Maronna, Martin and Yohai (2006, p. 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 $(T_j, \boldsymbol{C}_j)$ of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ for some constant $a > 0$, each with the same rate n^δ. If $(T_A, \boldsymbol{C}_A)$ is an estimator obtained by choosing one of the K estimators, then $(T_A, \boldsymbol{C}_A)$ is a consistent estimator of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ with rate n^δ by Pratt (1959). See Theorem 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, eg, n^δ where $0 < \delta \leq 0.5$, then the algorithm estimator is a consistent estimator of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ with the same rate as the attractors.

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

iv) Suppose the data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid and $P(\boldsymbol{x}_i = \boldsymbol{\mu}) < 1$. The elemental basic resampling algorithm estimator ($k = -1$) 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 of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$, and ii) follows from Pratt (1959). iii) Let $\gamma_{n,i}$ be the breakdown value of the ith attractor if the clean data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are in general position. The breakdown value γ_n of the algorithm estimator can be no lower than that of the worst attractor: $\gamma_n \geq \min(\gamma_{n,1}, ..., \gamma_{n,K}) \to 0.5$ as $n \to \infty$.

iv) Let $(\overline{\boldsymbol{x}}_{-1,j}, \boldsymbol{S}_{-1,j})$ be the classical estimator applied to a randomly drawn elemental set. Then $\overline{\boldsymbol{x}}_{-1,j}$ is the sample mean applied to $p+1$ iid cases. Hence $E[\overline{\boldsymbol{x}}_{-1,j}] = E(\boldsymbol{x}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\overline{\boldsymbol{x}}_{-1,j}) = \mathrm{Cov}(\boldsymbol{x})/(p+1) = \boldsymbol{\Sigma x}/(p+1)$ assuming second moments. So the $(\overline{\boldsymbol{x}}_{-1,j}, \boldsymbol{S}_{-1,j})$ are identically distributed

and inconsistent estimators of $(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$. Even without second moments, there exists $\epsilon > 0$ such that $P(\|\overline{\boldsymbol{x}}_{-1,j} - \boldsymbol{\mu}\| > \epsilon) = \delta_\epsilon > 0$ where the probability, ϵ and δ_ϵ do not depend on n since the distribution of $\overline{\boldsymbol{x}}_{-1,j}$ only depends on the distribution of the iid $\boldsymbol{x}_i$, not on n. Then $P(\min_j \|\overline{\boldsymbol{x}}_{-1,j} - \boldsymbol{\mu}\| > \epsilon) = P(\text{all } \|\overline{\boldsymbol{x}}_{-1,\text{j}} - \boldsymbol{\mu}\| > \epsilon) \to \delta_\epsilon^{\text{K}} > 0$ as $n \to \infty$ where equality would hold if the $\overline{\boldsymbol{x}}_{-1,j}$ were iid. Hence the "best start" that minimizes $\|\overline{\boldsymbol{x}}_{-1,j} - \boldsymbol{\mu}\|$ is inconsistent.

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

Since the FMCD estimator is a zero breakdown elemental concentration algorithm, the Hubert, Rousseeuw and Van Aelst (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 ∞.

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

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

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

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 $\boldsymbol{x}_i$ be random vectors for large sample theory, but the $\boldsymbol{x}_i$ 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* $(T_{0,j}, \boldsymbol{C}_{0,j})$. Each start is refined with k concentration steps, resulting in K_n *attractors* $(T_{k,j}, \boldsymbol{C}_{k,j})$, and the concentration attractor $(T_A, \boldsymbol{C}_A)$ 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, Gnanadesikan and Kettenring 1975, 1981) defined below is one example. Gnanadesikan and Kettenring (1972, p. 94–95) provide a similar algorithm. The DGK estimator is affine equivariant since the classical estimator is affine equivariant and Mahalanobis distances are invariant under affine transformations by Proposition 4.1. This section will show that the Olive (2004) MB estimator is 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) provides an estimator somewhat similar to the MB estimator.

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

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

The proof of the following theorem implies that a high breakdown estimator $(T, \boldsymbol{C})$ has $\text{MED}(D_i^2) \leq V$ and that the hyperellipsoid $\{\boldsymbol{x} | D_{\boldsymbol{x}}^2 \leq D_{(c_n)}^2\}$ that contains c_n 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, eg, $k = 10$. The theorem implies that the MB estimator $(T_{MB}, \boldsymbol{C}_{MB})$ is high breakdown.

Theorem 4.6. Suppose $(T, \boldsymbol{C})$ is a high breakdown estimator where $\boldsymbol{C}$ is a symmetric, positive definite $p \times p$ matrix if the contamination proportion

d_n/n is less than the breakdown value. Then the concentration attractor $(T_k, \boldsymbol{C}_k)$ is a high breakdown estimator if the coverage $c_n \approx n/2$ and the data are in general position.

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

$$0 < \|\boldsymbol{x}\|^2 \, \tau_n \leq \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \leq \|\boldsymbol{x}\|^2 \, \tau_1. \tag{4.6}$$

Let $\lambda_1 \geq \cdots \geq \lambda_n$ be the eigenvalues of $\boldsymbol{C}$. By (4.6),

$$\frac{1}{\lambda_1}\|\boldsymbol{x} - T\|^2 \leq (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) \leq \frac{1}{\lambda_n}\|\boldsymbol{x} - T\|^2. \tag{4.7}$$

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

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

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

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

To investigate the consistency and rate of robust estimators of multivariate location and dispersion, review Definition 3.16.

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 (2009, 2010) show 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 $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$.

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

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

Theorem 4.8, Lopuhaä (1999). a) If the start $(T, \boldsymbol{C})$ is inconsistent, then so is the attractor.

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

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

Let $\delta = 0.5$. Applying Theorem 4.8c) iteratively for a fixed number k of

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

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

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

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

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

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

$$= s^{-1}D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(n^{-\delta}). \tag{4.8}$$

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

$$b = D^2_\xi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{4.9}$$

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

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

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

be the population median of the population squared distances. By Remark 4.2, if $(T, \boldsymbol{C})$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ then $(T, \tilde{\boldsymbol{C}}) \equiv (T, D_{(c_n)}^2(T, \boldsymbol{C})\ \boldsymbol{C})$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, b\boldsymbol{\Sigma})$, and $D_i^2(T, \tilde{\boldsymbol{C}}) \leq 1$ is equivalent to $D_i^2(T, \boldsymbol{C}) \leq D_{(c_n)}^2(T, \boldsymbol{C}))$. Hence Lopuhaä (1999) theory applied to $(T, \tilde{\boldsymbol{C}})$ with $h = 1$ is equivalent to theory applied to the concentration estimator using the affine equivariant estimator $(T, \boldsymbol{C}) \equiv (T_{-1}, \boldsymbol{C}_{-1})$ as the start. Since b does not depend on s, concentration produces a sequence of estimators $(T_0, \boldsymbol{C}_0), ..., (T_k, \boldsymbol{C}_k)$ where $(T_j, \boldsymbol{C}_j)$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ where the constant $a > 0$ is the same for $j = 0, 1, ..., k$.

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

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

Proof. By Remark 4.1 the estimator is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with rate n^δ. By the remarks above, a will be the same for any consistent 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 Butler, Davies and Jhun (1993) and Cator and Lopuhaä (2009, 2010). If the MCD estimator is the start, then it is also the attractor by Rousseeuw and Van Driessen (1999) who show that concentration does not increase the MCD criterion. Hence $a = a_{MCD}$. □

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

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

$$\boldsymbol{C}_{FCH} = \frac{\text{MED}(D_i^2(T_A, \boldsymbol{C}_A))}{\chi^2_{p,0.5}} \boldsymbol{C}_A \tag{4.10}$$

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

Remark 4.3. The *MBA estimator* $(T_{MBA}, \boldsymbol{C}_{MBA})$ uses the attractor $(T_A, \boldsymbol{C}_A)$ with the smallest determinant. Hence the DGK estimator is used as the attractor if $det(\boldsymbol{C}_{DGK}) \leq det(\boldsymbol{C}_{MB})$, and the MB estimator is used as the attractor, otherwise. Then $T_{MBA} = T_A$ and $\boldsymbol{C}_{MBA}$ is computed using the right hand side of (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 MED($\boldsymbol{W}$) than half the data, then FCH uses the MB attractor. The FCH estimator only uses the attractor with the smallest determinant if $\|T_{DGK} - \text{MED}(\boldsymbol{W})\| \leq \text{MED}(D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p))$. Using the location criterion increases the outlier resistance of the FCH estimator for certain types of outliers, as will be seen in Section 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 $det(C_{FCH})$ is bounded away from 0 and ∞ if the data is in general position, even if nearly half of the cases are outliers.

Theorem 4.10. T_{FCH} is high breakdown if the clean data are in general position. Suppose (E1) holds. If $(T_A, \boldsymbol{C}_A)$ is the DGK or MB attractor

with the smallest determinant, then $(T_A, \boldsymbol{C}_A)$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$. Hence the MBA and FCH estimators are outlier resistant $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ where $c = u_{0.5}/\chi^2_{p,0.5}$, and $c = 1$ for multivariate normal data.

Proof. T_{FCH} is high breakdown since it is a bounded distance from MED($\boldsymbol{W}$) even if the number of outliers is close to $n/2$. Under (E1) the FCH and MBA estimators are asymptotically equivalent since $\|T_{DGK} - \mathrm{MED}(\boldsymbol{W})\| \to 0$ in probability. The estimator satisfies $0 < det(\boldsymbol{C}_{MCD}) \leq det(\boldsymbol{C}_A) \leq det(\boldsymbol{S}_{0,M}) < \infty$ 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 Theorem 4.9 since both starts are $\sqrt{n}$ consistent. Otherwise, the MB estimator $\boldsymbol{C}_{MB}$ is a biased estimator of $a_{MCD}\boldsymbol{\Sigma}$. But the DGK estimator $\boldsymbol{C}_{DGK}$ is a $\sqrt{n}$ consistent estimator of $a_{MCD}\boldsymbol{\Sigma}$ by Theorem 4.9 and $\|\boldsymbol{C}_{MCD} - \boldsymbol{C}_{DGK}\| = O_P(n^{-1/2})$. Thus the probability that the DGK attractor minimizes the determinant goes to one as $n \to \infty$, and $(T_A, \boldsymbol{C}_A)$ is asymptotically equivalent to the DGK estimator $(T_{DGK}, \boldsymbol{C}_{DGK})$.

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

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 $(T_{-1,j}, \boldsymbol{C}_{-1,j})$ that are affine equivariant $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, s_j\boldsymbol{\Sigma})$ where $s_j > 0$. The MCD criteria 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 Kth attractor is replaced by the MB estimator, and for $j < K$ the jth attractor $(T_{k,j}, \boldsymbol{C}_{k,j})$ is not used if $T_{k,j}$ has a greater Euclidean distance from MED($\boldsymbol{X}$) than half the data. Then the location estimator of the algorithm is high breakdown.

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

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

see Johnson and Wichern (1988, p. 103-104). The "MVE(c_n)" criterion is $h^p\sqrt{\det(\boldsymbol{S}_{\mathrm{k,j}})}$ where $h = D_{(c_n)}(\overline{\boldsymbol{x}}_{k,j}, \boldsymbol{S}_{k,j})$ (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 $D_i^2(T_{FCH}, \boldsymbol{C}_{FCH}) \leq \chi^2_{p,0.975}$, and let

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

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

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

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

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

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

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

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 $h^2 = \chi^2_{p,0.975}$ and $d = u_{0.5}/\chi^2_{p,q}$ where $q_2 \to q$ in probability as $n \to \infty$. Here $0.5 \leq q < 1$ depends on the elliptically contoured distribution, but $q = 0.5$ and $d = 1$ for multivariate normal data.

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

Simulations suggested $(T_{RMVN}, \boldsymbol{C}_{RMVN})$ gives useful estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for a variety of outlier configurations. Using 20 runs and $n = 1000$, the averages of the dispersion matrices were computed when the bulk of the data are iid $N_2(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = diag(1, 2)$. For clean data, FCH, RFCH and RMVN give $\sqrt{n}$ consistent estimators of $\boldsymbol{\Sigma}$, while 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 $diag(1.13, 1.85)$.

Next the data had $\gamma = 0.4$ and the outliers had $\boldsymbol{x} \sim N_2((0, 15)^T, 0.0001\boldsymbol{I}_2)$, a near point mass at the major axis. FCH, MB and RFCH estimated $2.6\boldsymbol{\Sigma}$ while RMVN estimated $\boldsymbol{\Sigma}$. FMCD and OGK failed to estimate d $\boldsymbol{\Sigma}$. Note

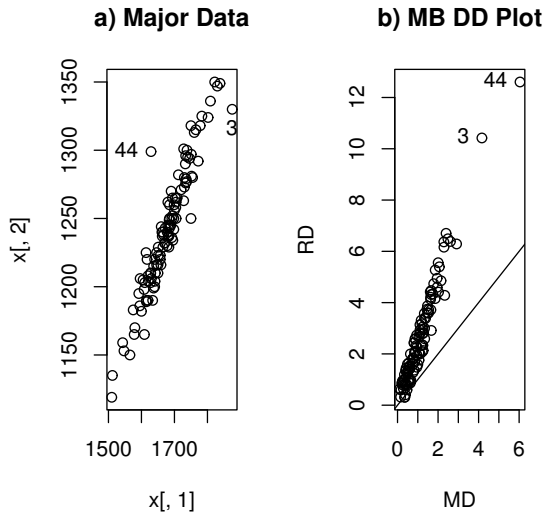

Figure 4.1: Plots for Major Data

that $\chi^2_{2,5/6}/\chi^2_{2,0.5} = 2.585$. 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 $\boldsymbol{x} \sim N_2((20,20)^T, \boldsymbol{\Sigma})$, a mean shift with the same covariance matrix as the clean cases. Rocke and Woodruff (1996) suggest that outliers with mean shift are hard to detect. FCH, FMCD, MB and RFCH estimated $2.6\boldsymbol{\Sigma}$ while RMVN estimated $\boldsymbol{\Sigma}$, and OGK failed. See Table 4.2. The *R command* is shown below.

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

Example 4.1. Tremearne (1911) recorded *height* = x[,1] and *height while kneeling* = 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 MED($\boldsymbol{W}$) = $(1680, 1240)^T$, but if the distances correspond to the contours of a covering ellipsoid, then case 44 has the largest distance. For $k = 0$, $(\overline{\boldsymbol{x}}_{0,M}, \boldsymbol{S}_{0,M})$ is the classical estimator applied to the "half set" of cases closest to MED($\boldsymbol{W}$) in Euclidean distance. The hypersphere (circle) centered at MED($\boldsymbol{W}$) that

covers half the data is small because the data density is high near MED($\boldsymbol{W}$). The median Euclidean distance is 59.661 and case 44 has Euclidean distance 77.987. Hence the intersection of the sphere and the data is a highly correlated clean ellipsoidal region. Figure 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 can not 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 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 $p = 2$.

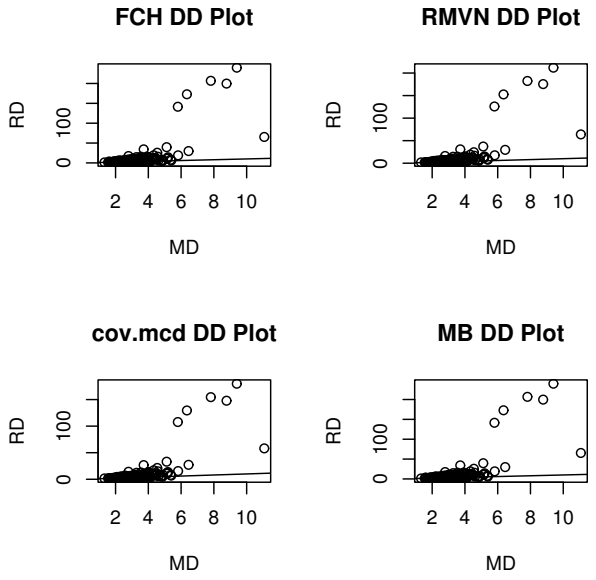

Figure 4.2: DD Plots for Gladstone Data

Example 4.2. The estimators can be useful when the data is not elliptically contoured. The Gladstone (1905-6) 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, 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

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

```
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
```

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

```
RMVN                                  FMCD
 0.988 -0.023 -0.007  0.021          0.227  -0.016  0.002  0.049
-0.023  1.964 -0.022 -0.002         -0.016   0.435 -0.014  0.0130
```

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

```
-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 55.648
```

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

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

If $W_{in} \sim N(0, \tau^2/n)$ for $i = 1, ..., r$ and if S_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, 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, eg $n^{1/3}$ rate, then the scaled sample variance $nS_W^2 \to \infty$ as $n \to \infty$.

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 $\boldsymbol{x} \sim N_p(\mathbf{0}, diag(1, ..., p))$. For the classical estimator, denoted by CLAS, $T_p = \overline{x}_p \sim N(0, p/n)$, and $nS^2(T_p) \approx p$

while $C_{p,p}$ is the sample variance of n iid $N(0,p)$ random variables. Hence $nS^2(C_{p,p}) \approx 2p^2$. RFCH, RMVN, FMCD and OGK use a "reweight for efficiency" concentration step that uses a random number of cases with percentage close to 97.5%. These four estimators had similar behavior. DGK, FCH and MB used about 50% of the cases and had similar behavior. By Lopuhaä (1999), estimators with less than $\sqrt{n}$ rate still have zero efficiency after the reweighting. Although FMCD, MB and OGK have not been proven to be $\sqrt{n}$ consistent, their values did not blow up even for $n = 5000$.

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

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

A simple simulation for outlier resistance is to count the number of times the minimum distance of the outliers is larger than the maximum distance of the clean cases. The simulation used 100 runs. If the count was 97, then in 97 data sets the outliers can be separated from the clean cases with a horizontal line in the DD plot, but in 3 data sets the robust distances did not achieve complete separation.

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

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

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

For near point mass outliers, an ellipsoid with very small volume can cover half of the data if the outliers are at one end of the ellipsoid 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 γ, as the near point mass is moved very far away from the bulk of the data, only the classical, MB and OGK estimators did not have numerical difficulties. Since the MCD estimator has smaller determinant than DGK while MVE has smaller volume than DGK, estimators like FMCD and MBA that use the MVE or MCD criterion without using location information will be vulnerable to these outliers. FMCD is also vulnerable to outliers if γ is slightly larger than γ_o given by (4.5).

Tables 4.4 and 4.5 help illustrate the results for the simulation. Large counts and small pm for fixed γ suggest greater ability to detect outliers. Values of p were 5, 10, 15, ..., 60. First consider the mean shift outliers and Table 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, although 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.

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

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

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 γ, OGK often has large bias for $c\boldsymbol{\Sigma}$. Then the outliers may need to be enormous before OGK can detect them. Also see Table 4.2, where OGK gave the outliers the largest distances for all runs, but $\boldsymbol{C}_{OGK}$ does not give a good estimate of $c\boldsymbol{\Sigma} = c\ \ diag(1, 2)$.

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

When p is increased to 8, 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 ellipsoid while the outliers fall in another hugely distant compact ellipsoid, then a concentration algorithm using a single start can sometimes tolerate nearly 25% outliers. For example, sup-

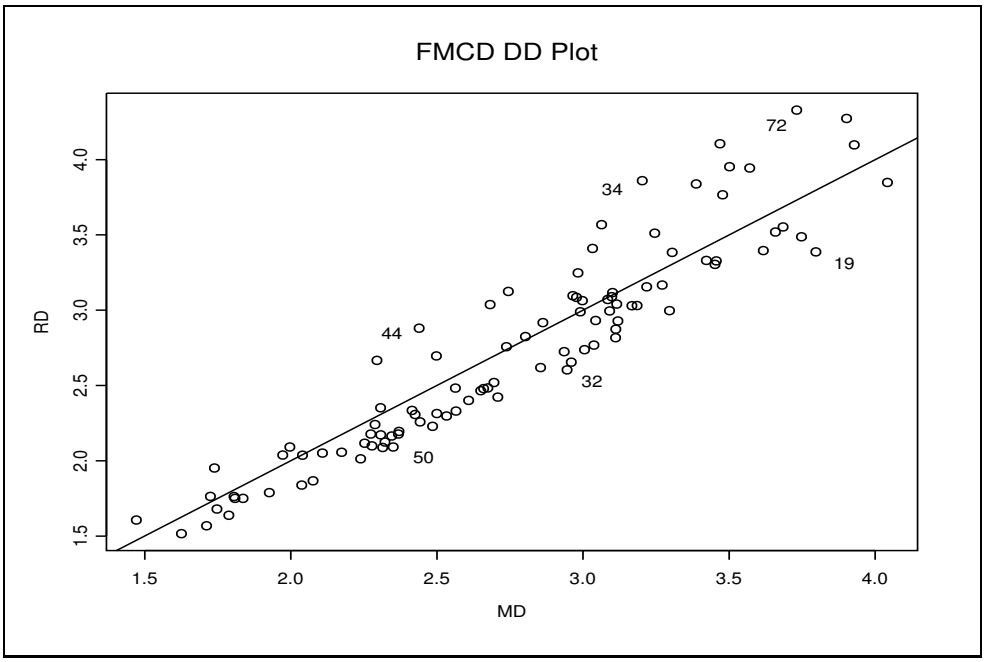

Figure 4.3: The FMCD Estimator Failed

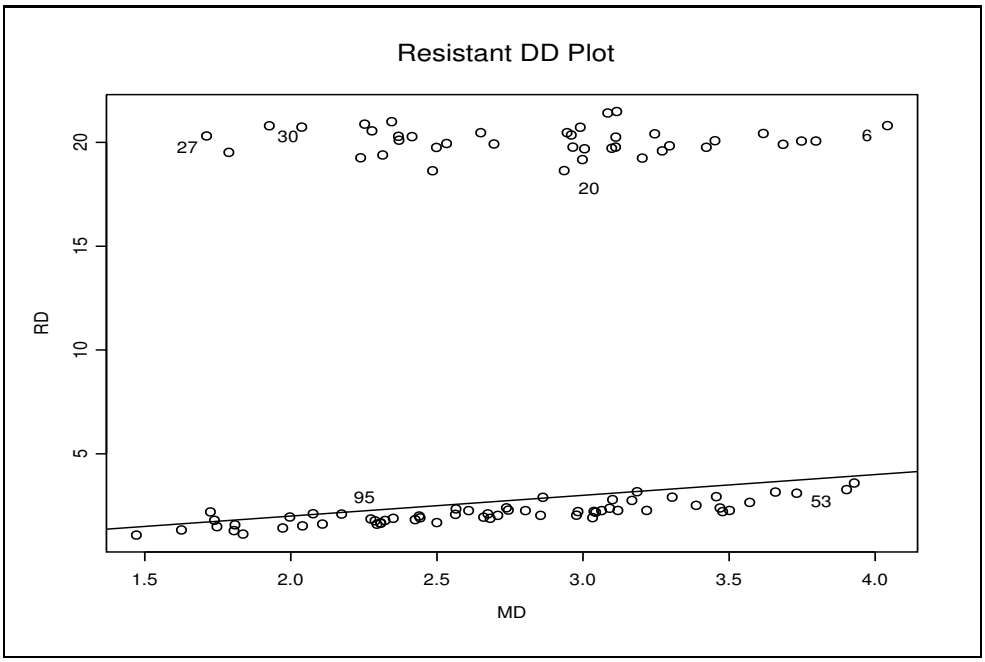

Figure 4.4: The Outliers are Large in the MBA DD Plot

pose that all $p+1$ cases in the elemental start are outliers but the covariance matrix is nonsingular so that the Mahalanobis distances can be computed. Then the classical estimator is applied to the $c_n \approx n/2$ cases with the smallest distances. Suppose the percentage of outliers is less than 25% and that all of the outliers are in this "half set." Then the sample mean applied to the c_n cases should be closer to the bulk of the data than to the cluster of outliers. Hence after a concentration step, the percentage of outliers will be reduced if the outliers are very far away. After the next concentration step the percentage of outliers will be further reduced and after several iterations, all c_n cases will be clean.

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

Proposition 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 $\boldsymbol{C}_A$ is less likely to be singular than the high breakdown MCD estimator $\boldsymbol{C}_{MCD}$.

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

Software

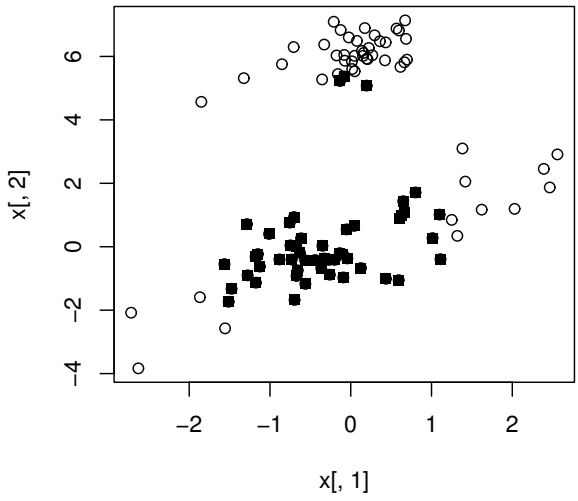

Figure 4.5: highlighted cases = half set with smallest RD = $(T_0, \boldsymbol{C}_0)$

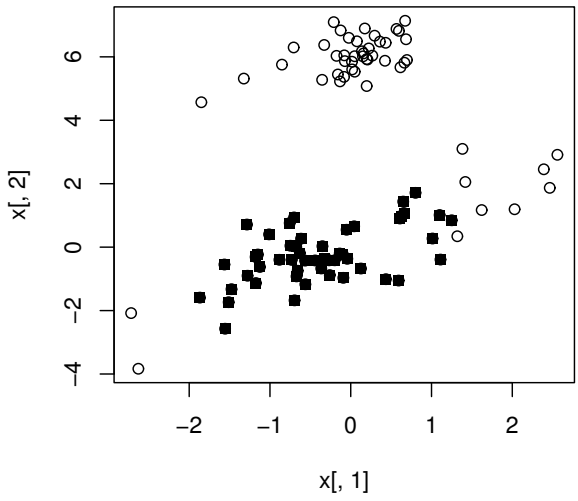

Figure 4.6: highlighted cases = half set with smallest RD = $(T_1, \boldsymbol{C}_1)$

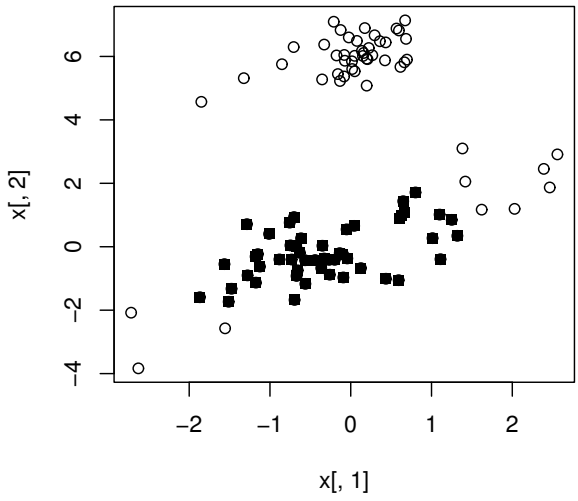

Figure 4.7: highlighted cases = half set with smallest RD = $(T_2, \boldsymbol{C}_2)$

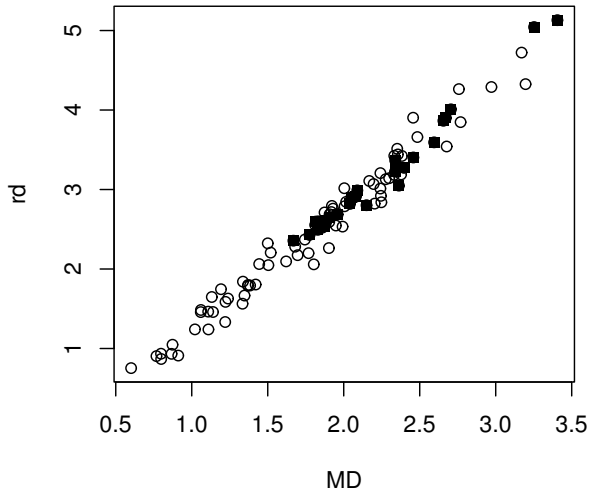

Figure 4.8: highlighted cases = outliers, RD = $(T_{0,D}, \boldsymbol{C}_{0,D})$

The `robustbase` library was downloaded from (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.

The `mpack` function
mldsim(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 `mldsim6` is similar but does not need the `library` command since it compares the FCH, RFCH, CMVE, RCMVE and MB estimators. 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.

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

The *concmv* function described in Problem 4.5 illustrates concentration where the start is $(\text{MED}(\boldsymbol{W}), diag([MAD(X_i)]^2))$. In Figures 4.5, 4.6, and 4.7, the highlighted cases are the half set with the smallest distances, and

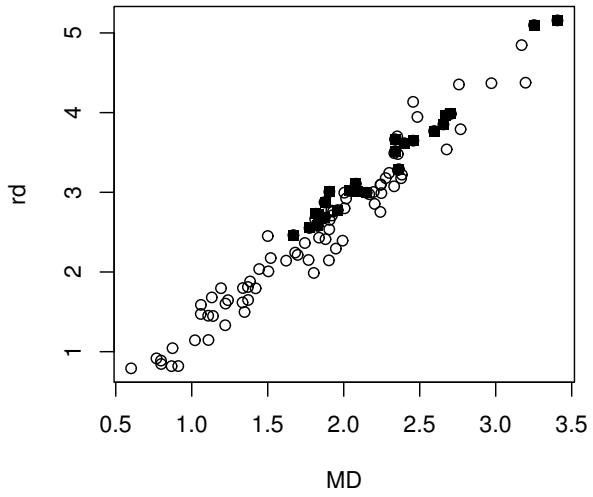

Figure 4.9: highlighted cases = outliers, RD = $(T_{1,D}, \boldsymbol{C}_{1,D})$

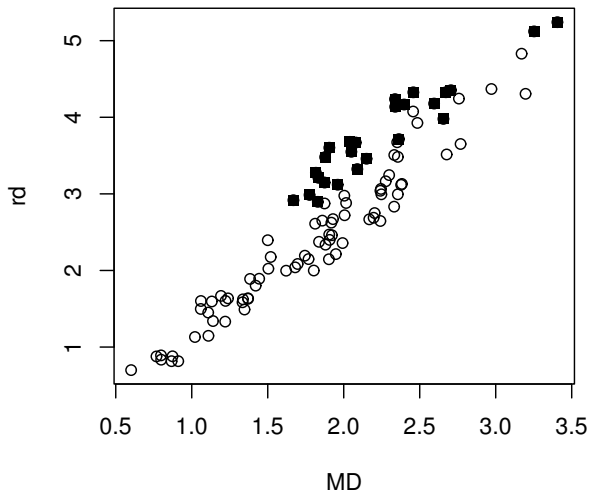

Figure 4.10: highlighted cases = outliers, RD = $(T_{2,D}, \boldsymbol{C}_{2,D})$

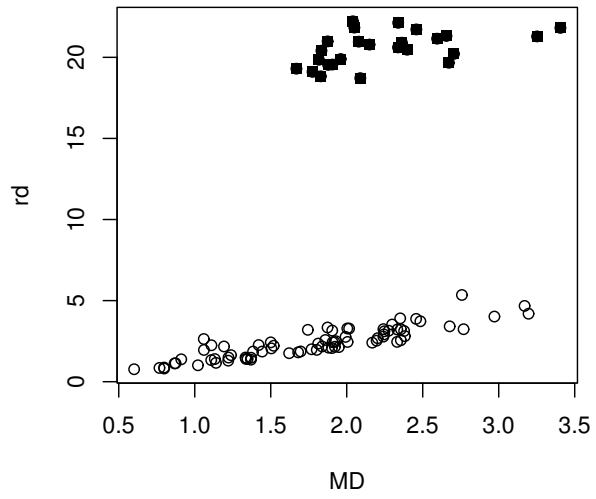

Figure 4.11: highlighted cases = outliers, RD = $(T_{3,D}, \boldsymbol{C}_{3,D})$

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 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 $\boldsymbol{x}_i \sim N_4(\boldsymbol{0}, diag(1, 2, 3, 4))$ while the outliers $\boldsymbol{x}_i \sim N_4((10, 10\sqrt{2}, 10\sqrt{3}, 20)^T,\ diag(1, 2, 3, 4))$.

4.6 Summary

1) Given a table of data $\boldsymbol{W}$ for variables $X_1, ..., X_p$, be able to find the **coordinatewise median** MED($\boldsymbol{W}$) and the **sample mean** $\overline{\boldsymbol{x}}$. If $\boldsymbol{x} =$

$(X_1, X_2, ..., X_p)^T$ where X_j corresponds to the jth column of $\boldsymbol{W}$, then $\text{MED}(\boldsymbol{W}) = (\text{MED}_{X_1}(n), ..., \text{MED}_{X_p}(n))^T$ where $\text{MED}_{X_j}(n) = \text{MED}(X_{j,1}, ..., X_{j,n})$ is the sample median of the data in the jth column. Similarly, $\overline{\boldsymbol{x}} = (\overline{X}_1, ..., \overline{X}_p)^T$ where $\overline{X}_j$ is the sample mean of the data in the jth column. See Q3.

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 (eg some plotted points are far from the bulk of the data or the plotted points follow two lines). See Q3.

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

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

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

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

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

the median ball, and the attractor with the smallest determinant, otherwise. Let $(T_A, \boldsymbol{C}_A)$ be the attractor used. Then the estimator $(T_{FCH}, \boldsymbol{C}_{FCH})$ takes $T_{FCH} = T_A$ and

$$\boldsymbol{C}_{FCH} = \frac{\mathrm{MED}(D_i^2(T_A, \boldsymbol{C}_A))}{\chi^2_{p,0.5}} \boldsymbol{C}_A \tag{4.12}$$

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

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

10) Brand name robust MLD estimators from the Rousseeuw and Yohai 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-τ, F-constrained-M and F-Stahel-Donoho are especially common.

4.7 Complements

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

The DGK estimator $(T_{k,D}, \boldsymbol{C}_{k,D})$ uses the classical estimator $(T_{-1,D}, \boldsymbol{C}_{-1,D})$ $= (\overline{\boldsymbol{x}}, \boldsymbol{S})$ as the only start. Thus $(\boldsymbol{\mu}_{-1,D}, \boldsymbol{\Sigma}_{-1,D})$ is the population mean and covariance matrix. For an elliptically contoured distribution with a nonsingular covariance matrix and for $t \geq 0$, $(\boldsymbol{\mu}_{t,D}, \boldsymbol{\Sigma}_{t,D})$ is the population mean and covariance matrix of the truncated distribution corresponding to the highest density region covering half the mass. Hence $\boldsymbol{\mu}_{t,D} = \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_{t,D} = c\boldsymbol{\Sigma}$

for some $c > 0$. Riani, Atkinson and Cerioli (2009) find the population mean and covariance matrices for such truncated multivariate normal distributions, using results from Tallis (1963).

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

Conjecture 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, $\boldsymbol{\mu}_{k,M} = \boldsymbol{\mu}$.

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

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 jth element of $\text{MED}(\boldsymbol{W})$ is the sample median of the jth predictor. Hence with $m-1$ outliers, $\text{MED}(\boldsymbol{W})$ can be driven to the "coordinatewise covering box" of the m clean cases. The boundaries of this box are at the min and

max of the clean cases from each predictor, and the lengths of the box edges equal the ranges R_i of the clean cases for the ith variable. If $d \approx m/2$ so that $\gamma \approx 1/3$, then the MED($\boldsymbol{W}$) can be moved to the boundary of the much smaller "coordinatewise IQR box" corresponding the 25th and 75th percentiles of the clean date. Then the edge lengths are approximately equal to the interquartile ranges of the clean cases.

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

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

One of the most effective methods for detecting outliers for large data sets or if $p > n$ is to use $D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$.

The MB estimator has outlier resistance similar to (MED($\boldsymbol{W}$), $\boldsymbol{I}_p$) for distant outliers but, as shown in Example 4.1, can be much more effective for detecting certain types of outliers that can not be found by marginal methods. For EC data, the MB estimator is best if the data is spherical about $\boldsymbol{\mu}$ or if the data is highly correlated with the major axis of the highest density region $\{\boldsymbol{x}_i : D_i^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leq d^2\}$.

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 $\boldsymbol{C}_{FCH}$ 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 $(T_{0,M}, \boldsymbol{C}_{0,M})$ that uses 50% trimming. $(T_{0,M}, \boldsymbol{C}_{0,M})$ is a high breakdown estimator by Corollary 4.7. Since only cases $\boldsymbol{x}_i$ such that $\|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\| \leq \text{MED}(\|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\|)$ are used, the largest eigenvalue of $\boldsymbol{C}_{0,50}$ is bounded if fewer than half of the cases are outliers by Lemma 4.3.

The geometric behavior of $(T_{0,M}, \boldsymbol{C}_{0,M})$ is simple. If the data $\boldsymbol{x}_i$ are MVN (or EC) then the highest density regions of the data are hyperellipsoids. The set of $\boldsymbol{x}$ closest to the coordinatewise median in Euclidean distance is a hypersphere. For EC data the highest density ellipsoid and hypersphere will have approximately the same center as the hypersphere, and the hypersphere

will be drawn towards the longest axis of the hyperellipsoid. Hence too much data will be trimmed in that direction. For example, if the data are MVN with $\boldsymbol{\Sigma} = \text{diag}(1, 2, ..., p)$ then $\boldsymbol{C}_{0,M}$ 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 $\boldsymbol{C}_{k,M}$ if the data is EC, and the determinant $|\boldsymbol{C}_{k,M}| < |\boldsymbol{C}_{0,M}|$ unless the attractor is equal $(T_{0,M}, \boldsymbol{C}_{0,M})$ by Proposition 4.4. The MB estimator $(T_{k,M}, \boldsymbol{C}_{k,M})$ 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) suggest an estimator similar to the MB estimator, also see Croux and Van Aelst (2002). Another estimator similar to MB was suggested by Wilk, Gnanadesikan, Huyett and Lauh (1962). See Gnanadesikan (1977, p. 134).

Recall that the *population squared Mahalanobis distance*

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

For elliptically contoured distributions, U has pdf given by (3.10), and the 50% highest density region has the form of the hyperellipsoid

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

where $U_{0.5}$ is the median of the distribution of U. For example, if the $\boldsymbol{x}$ are MVN, then U has the χ^2_p 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 $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$. See Theorem 4.9.

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

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

The theory for concentration algorithms is due to Hawkins and Olive (2002) and Olive and Hawkins (2010). The MBA estimator is due to Olive

(2004a). The computational and theoretical simplicity of the FCH estimator makes it one of the most useful robust estimators ever proposed. An important application of the robust algorithm estimators and of case diagnostics is to detect outliers. Sometimes it can be assumed that the analysis for influential cases and outliers was completely successful in classifying the cases into outliers and good or "clean" cases. Then classical procedures can be performed on the good cases. This assumption of perfect classification is often unreasonable, and it is useful to have robust procedures, such as the FCH estimator, that have rigorous asymptotic theory and are practical to compute. Since the FCH estimator is about an order of magnitude faster than alternative robust estimators, the FCH estimator may be useful for computationally intensive applications.

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

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

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

where $\mathrm{VAR}(\mathrm{X}) = \sigma^2(\mathrm{X})$ suggests that a robust estimator of dispersion can be created by replacing the sample standard deviation $\hat{\sigma}$ by a robust estimator of scale. Maronna and Zamar (2002) modify this idea to create a fairly fast (possibly high breakdown consistent) OGK estimator of multivariate location and dispersion. This estimator may be the leading competitor of the FCH estimator. Also see Alqallaf, Konis, Martin and Zamar (2002) and Mehrotra (1995). Woodruff and Rocke (1994) introduced the third technique, partitioning, which evaluates a start on a subset of the cases. Poor starts are discarded, and L of the best starts are evaluated on the entire data set. This idea is also used by Rocke and Woodruff (1996) and by Rousseeuw and Van Driessen (1999).

Billor, Hadi and Velleman (2000) have 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.

Croux, Dehon and Yadine (2010) claim that the practical Sign Covariance Matrix is high breakdown and that their practical k-step Spatial Sign Covariance Matrix is high breakdown and consistently estimates the orientation of the scatter matrix. The Sign Covariance Matix $\hat{\boldsymbol{\Sigma}}_S = \frac{1}{n}\sum_{i=1}^{n} \frac{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_n)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_n)^T}{\|\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_n\|^2}$ which is similar to the classical covariance estimator computed from $\boldsymbol{z}_i = \frac{\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_n}{\|\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_n\|}$. Here $\hat{\boldsymbol{\mu}}_n$ is the L_1-median or spatial median, defined as

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

is a fairly practical high breakdown estimator of multivariate location.

There certainly exist types of outlier configurations where the FMCD estimator outperforms the robust FCH estimator. The FCH estimators is vulnerable to outliers that lie inside the hypersphere based on the median Euclidean distance from the coordinatewise median. Although the FCH estimator should not be viewed as a replacement for the FMCD estimator, the FMCD estimator should be modified 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, FCH and classical estimators.

The simplest version of the MBA estimator only has two starts. A simple modification would be to add additional starts as in Problem 4.7. The Det-MCD estimator of Hubert, Rousseeuw, and Verdonck (2012) is very similar, uses 6 starts, but is not yet backed by theory.

Rousseeuw (1984) introduced the MCD and the minimum volume ellipsoid MVE(c_n) estimator. For the MVE estimator, $T(\boldsymbol{W})$ is the center of the minimum volume ellipsoid covering c_n of the observations and $\boldsymbol{C}(\boldsymbol{W})$ is determined from the same ellipsoid. T_{MVE} has a cube root rate and the limiting distribution is not Gaussian. See Davies (1992). Bernholdt and Fisher (2004) show that the MCD estimator can be computed with $O(n^v)$

complexity where $v = 1 + p(p+3)/2$ if $\boldsymbol{x}$ is a $p \times 1$ vector.

Rocke and Woodruff (1996, p. 1050) claim that any affine equivariant location and shape estimation method gives an unbiased location estimator and a shape estimator that has an expectation that is a multiple of the true shape for elliptically contoured distributions. Hence there are many candidate robust estimators of multivariate location and dispersion. See Cook, Hawkins and Weisberg (1993) for an exact algorithm for the MVE. Other papers on robust algorithms include Hawkins (1993, 1994), Hawkins and Olive (1999a), Hawkins and Simonoff (1993), He and Wang (1996), Olive (2004a), Olive and Hawkins (2007, 2008), Rousseeuw and Van Driessen (1999), Rousseeuw and van Zomeren (1990), Ruppert (1992), and Woodruff and Rocke (1993). Rousseeuw and Leroy (1987, § 7.1) also describes many methods.

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

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

Huber and Ronchetti (2009, p. 214, 233) note that theory for M-estimators of multivariate location and dispersion is "not entirely satisfactory with regard to joint estimation of" $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ 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 $(T_{MCD}, \boldsymbol{C}_{MCD})$ could denote the estimator from the exact algorithm while $(T_{AMCD}, \boldsymbol{C}_{AMCD})$ 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) point out that the AMVE is not the minimum volume ellipsoid (MVE) estimator.

Where the Rousseeuw-Yohai 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 ith case $\boldsymbol{x}_i$ be a $p \times 1$ random vector, and suppose the n cases are collected in an $n \times p$ matrix $\boldsymbol{W}$ with rows $\boldsymbol{x}_1^T, ..., \boldsymbol{x}_n^T$. 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, τ, projection based, constrained M, MM, and Stahel-Donoho estimators. **These estimators have computational complexity is higher than $O(n^p)$.** See Maronna, Martin and Yohai (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, Miller, and Wegman (2009) simulate the OGK and the Olive (2004a) median ball algorithm (MBA) estimators for $p = 100$ and n up to 50000, and note 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 roughly 100 to 200 times faster.

ii) No practical useful "high breakdown" estimator of multivariate location and dispersion from this paradigm has been shown to be consistent or 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 estimator has not been shown to be either high breakdown or consistent.** Also Hawkins and Olive (2002) showed that elemental concentration estimators using K starts are zero breakdown estimators. They are inconsistent if they use k concentration steps where k is fixed.

Papers with titles like Rousseeuw and Van Driessen (1999) "A Fast Algorithm for the Minimum Covariance Determinant Estimator" and Hubert, Rousseeuw and Van Aelst (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. Also see Olive (2012a).

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

If an estimator can not be computed in a reasonable amount of time, then most of its theoretical properties are only of academic interest (consistency of MCD is needed for the practical FCH estimator). What is of interest are the theoretical properties of the estimator actually used.

The central thesis of Hawkins and Olive (2002) was that, given the disconnect between the theoretically defined estimator and what can actually be computed, the theoretical properties of the former do not necessarily give useful guidance on the properties of the latter. Nearly all of the literature appears to overlook this disconnect, including Hubert, Rousseeuw and Van Aelst (2008) and Maronna, Martin and Yohai (2006).

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 $\boldsymbol{C}$ bounded away from 0 and ∞ 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, p. 200, 283) note 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 underlying 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 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.

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 mali-

ciously placed cases. In particular, the breakdown value of an estimator tells nothing about more important properties such as consistency or asymptotic normality.

4.8 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

R/Splus Problems

Use the command *source("G:/mpack.txt")* **to download the functions** and the command *source("G:/mrobdata.txt")* **to download the data. See Preface or Section 15.2.** Typing the name of the `mpack` function, eg *covmba*, will display the code for the function. Use the `args` command, eg *args(covmba)*, to display the needed arguments for the function.

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

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

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

4.2. Download the `rmaha` function that creates the robust Mahalanobis distances. Obtain outx2 as in Problem 4.1 b). *R* users need to enter the 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 in *R* the *library(MASS)* command), type the command

ddcomp(buxx). This will make 4 DD plots based on the DGK, FCH, FMCD and median ball estimators. The DGK and median ball estimators are the two attractors used by the FCH estimator. With the leftmost mouse button, move the cursor to 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 (1905-6) data and some infants are multivariate outliers.

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

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 MBA and FCH estimators can not always detect them. Type the command *concmv().* Hold the rightmost mouse button down (and in *R* click on *Stop*) to see the DD plot after one concentration step. The start uses the coordinatewise median and $diag([MAD(X_i)]^2)$. Repeat 4 more times to see the DD plot based on the attractor. The outliers have large values of X_2 and the highlighted cases have the smallest distances. Repeat the command *concmv()* several times. Sometimes the start will contain outliers but the attractor will be clean (none of the highlighted cases will be outliers), but sometimes concentration causes more and more of the highlighted cases to be outliers, so that the attractor is worse than the start. Copy one of the DD plots where none of the outliers are highlighted into *Word*.

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 4 more times to see the DD plot based on the attractor. In this problem, try to determine the proportion of outliers *gam* that the DGK estimator can detect for $p = 2, 4, 10$ and 20. Make a table of p and *gam*. For example the command

ddmv(p=2,gam=.4) suggests that the DGK estimator can tolerate nearly 40% outliers with $p = 2$, but the command *ddmv(p=4,gam=.4)* suggest that *gam* needs to be lowered (perhaps by 0.1 or 0.05). Try to make $0 < gam < 0.5$ as large as possible.

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 MED($\boldsymbol{x}$). 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.

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 $N_p(0, diag(1, ..., p))$.

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

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

4.9. The *R* function `cov.mcd` is a 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 (1905-6) 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.)