Statistics and Computing **12**: 191–200, 2002

Location adjustment for the minimum volume ellipsoid estimator

CHRISTOPHE CROUX*, GENTIANE HAESBROECK† and PETER J. ROUSSEEUW**

**Department of Applied Economics, Katholiek Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium*

†FEGSS, University of Liège, Bd du Rectorat 7, B-4000 Liège, Belgium

G.Haesbroeck@ulg.ac.be

***Department of Mathematics and Computer Science, Universiteit Antwerpen (U.I.A.), Universiteitsplein 1, B-2610 Antwerp, Belgium*

Received May 1998 and accepted July 2001

Estimating multivariate location and scatter with both affine equivariance and positive breakdown has always been difficult. A well-known estimator which satisfies both properties is the Minimum Volume Ellipsoid Estimator (MVE). Computing the exact MVE is often not feasible, so one usually resorts to an approximate algorithm. In the regression setup, algorithms for positive-breakdown estimators like Least Median of Squares typically recompute the intercept at each step, to improve the result. This approach is called *intercept adjustment*. In this paper we show that a similar technique, called *location adjustment*, can be applied to the MVE. For this purpose we use the Minimum Volume Ball (MVB), in order to lower the MVE objective function. An exact algorithm for calculating the MVB is presented. As an alternative to MVB location adjustment we propose L_1 *location adjustment*, which does not necessarily lower the MVE objective function but yields more efficient estimates for the location part. Simulations compare the two types of location adjustment. We also obtain the maxbias curves of both L_1 and the MVB in the multivariate setting, revealing the superiority of L_1.

Keywords: intercept adjustment, L_1 estimation, location estimation, location adjustment, minimum volume ellipsoid, robustness

1. Introduction

The Minimum Volume Ellipsoid (MVE) (Rousseeuw 1985) is defined as the smallest regular ellipsoid covering at least h elements of the data set $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^p$, where the MVE location estimator is the center of that ellipsoid and the MVE scatter estimator corresponds to its shape matrix. Equivalently, we can consider the minimization problem

$$(\hat{\mu}, \hat{S}) := \underset{\substack{(\mu,S)\in\mathbb{R}^p\times \mathrm{SPD}(p)\\ |S|=1}}{\operatorname{argmin}} d_h^2(\mu, S) \tag{1.1}$$

where $\mathrm{SPD}(p)$ is the set of all symmetric positive definite matrices $S \in \mathbb{R}^{p\times p}$. We call $\hat{S}$ the "shape matrix" because $\hat{S}$ determines the shape of the ellipsoid but not its magnitude, since necessarily $|\hat{S}| = 1$. By $d_h^2(\mu, S)$ we denote the hth ordered squared distance between $\mathbf{x}_i$ and μ in the metric given by S, i.e. $d_h^2(\mu, S) = \{(\mathbf{x}_i-\mu)^t S^{-1}(\mathbf{x}_i-\mu); 1 \le i \le n\}_{(h)} = \{\|\mathbf{x}_i-\mu\|_S^2; 1 \le i \le n\}_{(h)} = \text{median}_i \|\mathbf{x}_i - \mu\|_S^2$ where 'median' stands for the hth order statistic. Then the MVE estimator is given by

$$(\hat{\mu}, \hat{\Sigma}) := \left(\hat{\mu}, c(n, p, h) d_h^2(\hat{\mu}, \hat{S})\hat{S}\right) \tag{1.2}$$

where $c(n, p, h)$ is a correction factor to make $\hat{\Sigma}$ consistent for Σ at the normal model. If one wants to maximize the breakdown point of the estimator, the value of h in (1.1) and (1.2) can be set at $h = [\frac{n+p+1}{2}] \approx \frac{n}{2}$ (Lopuhaä and Rousseeuw 1991). But if (as is often the case) one knows that the fraction of outliers is at most α where $0 < \alpha < \frac{1}{2}$, we can work with the estimator MVE(α) in which $h = \lceil n(1-\alpha)\rceil$. The choice $\alpha = \frac{1}{4}$ is a good default value.

Throughout the paper we will assume not to be in the very degenerate situation where some h elements of $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^p$ all lie in a $(p-1)$-dimensional hyperplane. This is a *necessary* condition for the existence of the MVE. If such h points exist, they can be covered by ellipsoids with

arbitrary small volume, but the infimum volume cannot be attained by a regular ellipsoid. In (1.1) the infimum of $d_h^2(\mu, S)$ also becomes zero but again cannot be attained. The condition is also *sufficient* for the existence of the MVE, since the infimum volume is then the minimum of a finite number $\binom{n}{h}$ of strictly positive volumes.

In a regression model $y_i = \beta^t \mathbf{x}_i + \alpha + \varepsilon_i (i = 1, \ldots, n)$ with slope parameter $\beta \in \mathbb{R}^{p-1}$ and intercept parameter $\alpha \in \mathbb{R}$, the Least Median of Squares (LMS) estimator (Rousseeuw 1984) of (α, β) is defined by

$$(\hat{\alpha}, \hat{\beta}) = \underset{(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^{p-1}}{\operatorname{argmin}} \underset{i}{\operatorname{median}} (y_i - \beta^t \mathbf{x}_i - \alpha)^2. \tag{1.3}$$

Usually, the intercept estimate $\hat{\alpha}$ is computed conditionally on the value of the slope estimate $\hat{\beta}$ in order to lower the value of the objective function. This intercept adjustment process consists of splitting up the minimization problem (1.3) into two parts:

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^{p-1}}{\operatorname{argmin}} \underset{i}{\operatorname{median}} (y_i - \beta^t \mathbf{x}_i - \hat{\alpha}(\beta))^2$$

with

$$\hat{\alpha}(\beta) := \underset{\alpha\in\mathbb{R}}{\operatorname{argmin}} \underset{i}{\operatorname{median}} (y_i - \beta^t \mathbf{x}_i - \alpha)^2.$$

This corresponds to saying that $\hat{\alpha}(\beta)$ is the univariate LMS location estimate of the n numbers $y_i - \beta^t \mathbf{x}_i \, (i = 1, \ldots, n)$. Fortunately there exists an exact algorithm for the univariate LMS location estimate, since it is the midpoint of the shortest interval containing h observations (Rousseeuw 1984, Theorem 2).

Analogously, we can rewrite (1.1) as follows:

$$\hat{S} = \underset{\substack{S\in \mathrm{SPD}(p)\\ |S|=1}}{\operatorname{argmin}} d_h^2(\hat{\mu}(S), S) \tag{1.4}$$

where for a given S with $|S| = 1$ we put

$$\begin{aligned} \hat{\mu}(S) &:= \underset{\mu\in\mathbb{R}^p}{\operatorname{argmin}} d_h(\mu, S) \\ &= \underset{\mu\in\mathbb{R}^p}{\operatorname{argmin}} \underset{i}{\operatorname{median}} \|\mathbf{x}_i - \mu\|_S \\ &= \underset{\mu\in\mathbb{R}^p}{\operatorname{argmin}} \underset{i}{\operatorname{median}} \|S^{-1/2}\mathbf{x}_i - S^{-1/2}\mu\| \end{aligned} \tag{1.5}$$

where $S^{1/2}$ denotes the symmetric root of S (i.e. $S^{1/2}S^{1/2} = S$ with $S^{1/2}$ symmetric). Using the transformed data set $Y = \{\mathbf{y}_i = S^{-1/2}\mathbf{x}_i, i = 1, \ldots, n\}$, we obtain

$$\hat{\mu}(S) = S^{1/2} \underset{\theta}{\operatorname{argmin}} \underset{i}{\operatorname{median}} \|\mathbf{y}_i - \theta\|. \tag{1.6}$$

(Note that $\hat{\mu}(\hat{S})$ with $\hat{S}$ defined by (1.4), equals $\hat{\mu}$, and therefore remains affine equivariant.) The value of θ minimizing $\operatorname{median}_i \|\mathbf{y}_i - \theta\|$ is the center of the ball with smallest (nonzero) volume that covers at least h points of the data set Y. This corresponds to the minimum volume ball estimator defined by Rousseeuw (1984, p. 877). It is known that this estimator is orthogonal equivariant and has a 50% breakdown point. However, to our knowledge, no exact algorithm to compute this estimator has yet appeared in the literature. Section 2 of this paper presents such an algorithm. The $(p + 1)$-subset algorithm with location adjustment for computing the MVE is outlined in Section 3.

The purpose of location adjustment using the MVB is to lower the value of the objective function (1.1). It is similar to the well known intercept adjustment technique in robust regression which is believed to be beneficial. In Section 4 we show by means of a simulation study that location and intercept adjustments indeed lower the value of the objective function, but that the statistical benefit of this decrease it not so important as it might seem. On the other hand, location adjustment using the L_1-location estimator (presented in Section 5) does increase the statistical efficiency of the location estimator, while not lowering the value of the objective function. Expressions for the maxbias curves of the MVB and the L_1-estimator are derived in Section 6, allowing for a better theoretical understanding of the robustness behavior of these two orthogonally equivariant estimators. A comparison with the Feasible Solution Algorithm of Hawkins (1993b) is made in Section 7, while Section 8 concludes.

2. Computation of the minimum volume ball

Given $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\} \subset \mathbb{R}^p$, the MVB location estimator is defined as

$$\mathrm{MVB}(Y) = \underset{\mu\in\mathbb{R}^p}{\operatorname{argmin}} \underset{i}{\operatorname{median}} \|\mathbf{y}_i - \mu\| \tag{2.1}$$

where again the median stands for the hth order statistic.

The pseudocode below gives an *exact algorithm* for the MVB estimator:

1. Initialize R_{best} by $+\infty$.
2. For any integer $2 \le k \le p + 1$ and any k-subset $J = \{i_1, \ldots, i_k\} \subset \{1, \ldots, n\}$ do:
   - 2.1 Put $A_J :=$ affinespan $\{\mathbf{y}_j : j \in J\}$. If $\dim(A_J) < k - 1$, goto 2 (i.e., drop this J).
   - 2.2 Therefore, $\dim(A_J) = k - 1$. Determine the unique point μ_J in A_J that lies at the same Euclidean distance of all $\mathbf{y}_j$ for $j \in J$ by solving a $k \times k$ linear system of equations.
   - 2.3 Compute $R_J := \operatorname{median}_i \|\mathbf{y}_i - \mu_J\|$. If $R_J \ge R_{\text{best}}$ goto 2.
   - 2.4 Put $\mu_{\text{best}} := \mu_J$ and $R_{\text{best}} := R_J$.
3. Report $\mathrm{MVB}(Y) := \mu_{\text{best}}$ as well as R_{best}.

Theorem 1. *This algorithm yields the exact MVB estimator* (2.1).

The proof is given in the Appendix.

The problem is to find k and the optimal subset $\{i_1, \ldots, i_k\}$. Going through all possible subsets J of all possible sizes, and computing R_J in $O(n)$ time for each of them, yields the complexity $O((n^2 + \cdots + n^{p+1})n) = O(n^{p+2})$. Although the exact MVB takes a lot of time, its number of operations is still polynomial in n.

Since the exact MVB algorithm is too time-consuming, we also consider an approximate algorithm where step 2 only draws *Nsamp* subsets of size $p+1$. (By simulations, we found that the subset J yielding R_{best} and μ_{best} has size $(p+1)$ with a fairly high probability.)

Another alternative is a very rough 1-subset approximate algorithm:

1. For each $j = 1, \ldots, n$, compute $R_j := \text{median}_i \|\mathbf{y}_i - \mathbf{y}_j\|$.
2. Set MVB(Y) equal to the observation $\mathbf{y}_j$ attaining the lowest R_j.

This algorithm is less precise but it remains orthogonal equivariant and still has a 50% breakdown point. Moreover, it only takes $O(n^2)$ operations.

3. Location adjustment by the MVB

Finding the exact solution of the MVE minimization problem (1.1) is often not feasible. Therefore, one usually resorts to the approximate $(p+1)$-subset algorithm. One can easily adapt this algorithm to incorporate a location adjustment using the minimum volume ball estimator. This leads to the following algorithm:

1. Initialize R_{best} by $+\infty$.
2. For any $(p+1)$-subset $J \subset \{1, \ldots, n\}$ do:
   - 2.1 Compute $\mu_J = \frac{1}{p+1} \sum_{i \in J} \mathbf{x}_i$ and $C_J = \frac{1}{p} \sum_{i \in J} (\mathbf{x}_i - \mu_J)(\mathbf{x}_i - \mu_J)^t$. If $|C_J| = 0$ goto 2.
   - 2.2 Compute $S_J = |C_J|^{-1/p} C_J$ hence $|S_J| = 1$.
   - 2.3 Transform the data set X to $Y = \{\mathbf{y}_i = S_J^{-1/2} \mathbf{x}_i; i = 1, \ldots, n\}$.
   - 2.4 Compute the estimate $\theta_J := \text{MVB}(Y)$ and put $R_J := \text{median}_i \|\mathbf{y}_i - \theta_J\|$.
   - 2.5 If $R_J \geq R_{\text{best}}$ goto 2.
   - 2.6 Put $S_{\text{best}} := S_J$, $\theta_{\text{best}} := \theta_J$ and $R_{\text{best}} := R_J$.
3. Report the final estimate $(\hat{\mu}, \hat{\Sigma})$ where $\hat{\mu} = S_{\text{best}}^{1/2} \theta_{\text{best}}$ and $\hat{\Sigma} = c(n, p, h) R_{\text{best}}^2 S_{\text{best}}$. Note that the minimized objective value (1.1) equals R_{best}^2.

As the dimension p increases, it becomes infeasible to consider all $\binom{n}{p+1}$ subsets. Then we can still search over a random selection of *Nsamp* subsets of size $p+1$. One can also apply the MVB adjustment only once, as a final improvement to the usual $(p+1)$-subset algorithm for the MVE.

Note that all these versions of the MVE combined with MVB location adjustment are affine equivariant methods, because we apply the orthogonally equivariant MVB to the data in the MVE metric.

4. Simulations

The $(p+1)$-subset algorithm for the MVE was described by Rousseeuw and Leroy (1987, pp. 259–260). An actual program was provided by Rousseeuw and van Zomeren (1990) and incorporated in S-Plus and SAS. It is also easy to implement the algorithm in a matrix language like Gauss, which we did here. We compared the following estimators of location and scatter: the original $(p+1)$-subset estimator $(\hat{\mu}^p, \hat{\Sigma}^p)$, the $(p+1)$-subset estimator $(\tilde{\mu}, \tilde{\Sigma})$ with MVB adjustment at each step, and the $(p+1)$-subset estimator $(\tilde{\tilde{\mu}}, \tilde{\tilde{\Sigma}})$ with a single MVB adjustment at the end.

We generated two types of data configurations. The first one is the normal situation where $X_i \sim N(0, \mathbf{I}_p)$ for $i = 1, \ldots, n$. In the second situation, 20% of the observations are contaminated by replacing them by $100e_1$ where e_1 is the first unit vector. This yields a cluster of extreme outliers. We considered $p = 2$, $n = 30$ with *Nsamp* $= 400$, and $p = 3$, $n = 40$ with *Nsamp* $= 500$.

Summary values over $m = 500$ runs were computed, such as the bias and the mean squared error of the location estimators

$$\text{Bias}(\hat{\mu}) = \|\bar{\mu} - \mu\| = \left\| \left(\frac{1}{m} \sum_{k=1}^{m} \hat{\mu}^k \right) - \mu \right\| \tag{4.1}$$

$$\text{MSE}(\hat{\mu}) = \frac{1}{m} \sum_{k=1}^{m} \|\hat{\mu}^k - \mu\|^2, \tag{4.2}$$

where $\hat{\mu}^k$ is the estimate of location from the kth simulated sample and the true parameter is $\mu = 0$. To measure the deviation from sphericity of the estimated scatter matrix $\hat{\Sigma}^k$ of the kth sample, we calculated

$$\phi_k = \frac{\text{trace}(\hat{\Sigma}^k / p)^p}{\det(\hat{\Sigma}^k)}$$

according to Maronna and Yohai (1995). In Table 1, we reported

$$\underset{k=1,\ldots,m}{\text{median}} \ln \phi_k.$$

Note that the matrices $\hat{\Sigma}^p$ and $\tilde{\tilde{\Sigma}}$ only differ by a factor, and therefore have the same deviation from sphericity. Finally, the average value of the objective function (1.1) over the m runs is listed.

In Table 1, we see that applying the MVB adjustment does lower the MVE objective function compared to the original $(p+1)$-subset MVE algorithm, especially when the adjustment is carried out in each step. Indeed, by construction of the algorithms we know that $(\tilde{\mu}, \tilde{\Sigma})$ always yields lower values for the objective function than $(\tilde{\tilde{\mu}}, \tilde{\tilde{\Sigma}})$, which on its turn yields lower values than $(\hat{\mu}^p, \hat{\Sigma}^p)$. On the other hand, the adjustments don't have much impact on the bias, MSE and $\text{median}_k \ln \phi_k$ since these do not improve much. Even when the adjustment is carried out in every step, we do not gain much statistical precision, while the computation time increases drastically. There are however some cases (e.g. $p = 3$ for Normal data, ...) where the improvement turns out to be significant.

These results are similar to the regression framework. For instance, Table 2 compares the Least Trimmed Squares (LTS) estimator computed with or without intercept adjustment. The univariate LTS can be computed exactly with an algorithm of Rousseeuw and Leroy (1987, pp. 171–172). Let us now generate

Table 1. *Using location adjustment by the MVB*

		Normal data			20% contaminated data		
		$(\hat{\mu}^p, \hat{\Sigma}^p)$	$(\tilde{\mu}, \tilde{\Sigma})$	$(\tilde{\tilde{\mu}}, \tilde{\tilde{\Sigma}})$	$(\hat{\mu}^p, \hat{\Sigma}^p)$	$(\tilde{\mu}, \tilde{\Sigma})$	$(\tilde{\tilde{\mu}}, \tilde{\tilde{\Sigma}})$
$p = 2$	Bias($\hat{\mu}$)	0.008	0.008	0.008	0.028	0.015	0.018
$n = 30$	MSE($\hat{\mu}$)	0.234	0.229	0.231	0.259	0.228	0.252
	$\text{med}_k \ln \phi_k$	0.594	0.586	0.594	0.408	0.364	0.408
	Ave_kObj_k	1.016	0.878	0.985	1.508	1.293	1.453
$p = 3$	Bias($\hat{\mu}$)	0.018	0.023	0.023	0.037	0.035	0.028
$n = 40$	MSE($\hat{\mu}$)	0.295	0.266	0.319	0.310	0.301	0.353
	$\text{med}_k \ln \phi_k$	0.921	0.786	0.921	0.702	0.719	0.702
	Ave_kObj_k	2.150	1.897	2.108	2.991	2.616	2.885

In each of four situations, simulating $m = 500$ samples yields the bias and MSE of the location estimators, the $\text{median}_{k=1,\ldots,m} \ln \phi_k$ of the scatter estimators, and the average value of the MVE objective function. The standard errors around the reported values are about 0.01 for the bias and the mean squared error. The standard errors for the average value of the objective function are between 0.01 and 0.03. For the median of the deviations of the sphericity measures the standard error is of the order 0.025.

Table 2. *Using intercept adjustment when computing LTS regression*

	Standard data			20% vertical outliers			20% horizontal outliers		
	$(\hat{\beta}^p, \hat{\alpha}^p)$	$(\tilde{\beta}, \tilde{\alpha})$	$(\tilde{\tilde{\beta}}, \tilde{\tilde{\alpha}})$	$(\hat{\beta}^p, \hat{\alpha}^p)$	$(\tilde{\beta}, \tilde{\alpha})$	$(\tilde{\tilde{\beta}}, \tilde{\tilde{\alpha}})$	$(\hat{\beta}^p, \hat{\alpha}^p)$	$(\tilde{\beta}, \tilde{\alpha})$	$(\tilde{\tilde{\beta}}, \tilde{\tilde{\alpha}})$
$100 \times \text{Bias}(\hat{\beta})$	0.142	0.079	0.142	0.331	0.262	0.331	0.125	0.172	0.125
$100 \times \text{MSE}(\hat{\beta})$	0.463	0.460	0.463	0.422	0.430	0.422	0.414	0.385	0.414
$100 \times \text{Bias}(\hat{\alpha})$	1.677	1.749	1.700	0.227	0.275	0.786	2.010	1.428	2.720
$\text{MSE}(\hat{\alpha})$	0.129	0.124	0.126	0.134	0.123	0.122	0.020	0.014	0.027
Ave_kObj_k	0.810	0.781	0.799	1.195	1.148	1.174	1.194	1.147	1.174

In three situations, simulating $m = 500$ samples gives the bias and MSE of the slope and intercept estimators and the average value of the LTS objective function. Standard errors around the simulated bias are about 0.1 for $100\times$ the regression slope vector and 0.01 for $100\times$ the intercept. For the MSE we have standard errors of about 0.15 for $100\times$ the MSE of the slope vector and 0.01 for the MSE of the intercept.

three different situations. In the first one, the model is given by

$$y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \alpha + e_i \quad \text{for } i = 1, \ldots, 40 \tag{4.3}$$

with $\beta_1 = \beta_2 = \beta_3 = \alpha = 1$. Here, $e_i \sim N(0, 1)$ and the explanatory variables are generated independently as $x_{i,j} \sim N(0, 10)$ for $j = 1, \ldots, 3$. The second configuration replaces the first 8 points by outliers in the y-direction: $e_i \sim N(10, 1)$ for $i = 1, \ldots, 8$. In the third situation, these were replaced by outliers in the x-direction with $x_{i,1} \sim N(100, 10)$ and $e_i \sim N(0, 1)$ for $i = 1, \ldots, 8$.

Bias and MSE, defined as in (4.1) and (4.2), were computed for estimators of the regression and intercept parameters using the p-subset algorithm without intercept adjustment, with intercept adjustment in every step, and using intercept adjustment only at the final stage. In Table 2 we see that intercept adjustment indeed lowers the LTS objective function, especially if the adjustment is carried out at each step, whereas the bias and MSE of the coefficients do not change significantly. This matches our simulation results for multivariate location and scatter (Table 1).

5. L_1 adjustment

We have seen that MVB adjustment lowers the value of the MVE objective function. But MVB adjustment does not appreciably increase the finite sample efficiency of the estimator, while it requires much computation time. That is why we thought of replacing the MVB adjustment by a different adjustment which is less time consuming. The L_1 location estimator appears suitable since it has the same breakdown and equivariance properties as the MVB estimator, and is easier to compute. Moreover, it has a better statistical efficiency than the MVB. For a given p-dimensional data set $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ the L_1 estimator $\mu_L(Y)$ is the solution of the minimization problem

$$\mu_L(Y) = \underset{\mu \in \mathbb{R}^p}{\text{argmin}} \sum_{i=1}^{n} \|\mathbf{y}_i - \mu\|. \tag{5.1}$$

A fast algorithm for the L_1 estimator is given in Hössjer and Croux (1995).

We can now carry out location adjustment by means of the L_1 estimator. For this it suffices to take the algorithm in Section 3, and to replace step 2.4 by

Table 3. *Using L_1 adjustment when computing the MVE*

		Normal data			20% contaminated data		
		$(\hat{\mu}^p, \hat{\Sigma}^p)$	$(\tilde{\mu}, \tilde{\Sigma})$	$(\tilde{\tilde{\mu}}, \tilde{\tilde{\Sigma}})$	$(\hat{\mu}^p, \hat{\Sigma}^p)$	$(\tilde{\mu}, \tilde{\Sigma})$	$(\tilde{\tilde{\mu}}, \tilde{\tilde{\Sigma}})$
$p = 2$	Bias	0.0079	0.0062	0.0034	0.0211	0.4082	0.4015
$n = 30$	MSE	0.234	0.093	0.094	0.259	0.277	0.272
	$\text{med}_k \ln \phi_k$	0.594	0.531	0.594	0.408	0.404	0.408
	$\text{Ave}_k \text{Obj}_k$	1.016	1.055	1.332	1.508	1.677	2.178
$p = 3$	Bias	0.0178	0.0036	0.0057	0.0369	0.4605	0.4601
$n = 40$	MSE	0.295	0.092	0.093	0.310	0.328	0.328
	$\text{med}_k \ln \phi_k$	0.921	0.845	0.921	0.702	0.693	0.702
	$\text{Ave}_k \text{Obj}_k$	2.150	2.093	2.461	2.991	3.109	3.695

In each of four situations, simulating $m = 500$ samples yields the bias and MSE of the location estimators, the $\text{median}_{k=1,\ldots,m} \ln \phi_k$ of the scatter estimators, and the average value of the MVE objective function.

2.4' Compute the L_1 estimate $\theta_J := \mu_L(Y)$ and calculate $R_J := \text{median}_i \|\mathbf{y}_i - \theta_J\|$.

Everything else remains the same. Of course, we can do this both in the case of exhaustive search over all $(p+1)$-subsets as in the version of *Nsamp* randomly drawn $(p+1)$-subsets. Also, we can apply the L_1 location adjustment only once at the end. This corresponds to the two-stage estimator defined as:

compute the MVE estimator $(\hat{\mu}, \hat{\Sigma})$ and then replace $\hat{\mu}$ by

$$\tilde{\tilde{\mu}} := \underset{\mu \in \mathbb{R}^p}{\text{argmin}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mu\|_{\hat{\Sigma}}. \tag{5.2}$$

The breakdown point and asymptotic properties of this estimator were studied in Lopuhaä (1992), Hössjer and Croux (1995). Note that all these versions of "MVE with L_1 location adjustment" remain exactly affine equivariant.

We performed a modest simulation study to compare the random $(p+1)$-subset algorithm for MVE computed with and without the L_1 location adjustment. The simulation setup is similar to the one described in Section 4. Table 3 summarizes the results. As expected, this type of location adjustment does not reduce the MVE objective function. Surprisingly, the effect of L_1 adjustment on $\text{Bias}(\hat{\mu})$, $\text{MSE}(\hat{\mu})$ and $\text{median}_k \ln \phi_k$ is the same whether the adjustment is carried out at each step or only at the end. The latter version, denoted as $(\tilde{\tilde{\mu}}, \tilde{\tilde{\Sigma}})$, is of course the fastest to compute.

So, is L_1 adjustment beneficial? For the scatter matrix estimator, we see that $\text{med}_k \ln \phi_k$ remains about the same in all situations. For the location estimates, it depends on whether the data are normal or contaminated. For normal (uncontaminated) data, the L_1 adjustment preserves the small bias and substantially reduces the MSE, since L_1 has a better statistical efficiency. But for contaminated data the MSE remains the same, whereas the bias becomes much higher.

6. Maxbias curves

In this section, we will compare the maxbias curves of the location estimators MVB and L_1 in the multivariate setting. The maxbias of a location estimator T at the model distribution F and a given amount of contamination ε is given by

$$B(\varepsilon, T, F) = \sup_H \|T((1-\varepsilon)F + \varepsilon H) - T(F)\| \tag{6.1}$$

where H can be any distribution.

We will compute the maxbias curves for point contamination at normal model distributions, and due to orthogonal and translation equivariance we can take $F = N(\mathbf{0}, \Sigma)$ with $\Sigma = \text{diag}(\lambda_1, \ldots, \lambda_p)$, $\lambda_1 \geq \cdots \geq \lambda_p > 0$. The univariate standard normal distribution function will be denoted by Φ. All proofs are given in the Appendix.

Theorem 2. (a) *The maxbias $B(\varepsilon, L_1, F)$ of the L_1 location estimator at F is given by the solution b of the equation*

$$\frac{1}{\sqrt{2\pi\lambda_1}} \int_{-\infty}^{+\infty} \int_{0}^{+\infty} \frac{z}{\sqrt{z^2+v}} e^{-\frac{1}{2}\left(\frac{z+b}{\sqrt{\lambda_1}}\right)^2} g_V(v)\, dv\, dz = -\frac{\varepsilon}{1-\varepsilon} \tag{6.2}$$

where g_V denotes the probability density function of the random variable $V = \sum_{i=2}^{p} \lambda_i Y_i^2$ where the variables Y_i are i.i.d. univariate standard normal.

(b) *The maxbias $B(\varepsilon, MVB, F)$ is the positive solution b of the equation*

$$\int_{\|y\| \leq R_\varepsilon^+} g((y + be_1)^t \Sigma^{-1} (y + be_1))\, dy = \frac{\sqrt{\lambda_1 \ldots \lambda_p}(1-2\varepsilon)}{2(1-\varepsilon)}, \tag{6.3}$$

where R_ε^+ is defined implicitly by

$$\int_{\|y\| \leq R_\varepsilon^+} g(y^t \Sigma^{-1} y)\, dy = \frac{\sqrt{\lambda_1 \ldots \lambda_p}}{2(1-\varepsilon)} \tag{6.4}$$

where $g(t) = (\frac{1}{\sqrt{2\pi}})^p e^{-\frac{t}{2}}$ and e_1 is the first unit vector.

(c) *For the maxbias of translation equivariant multivariate location estimators at normal models, the lower bound of He*

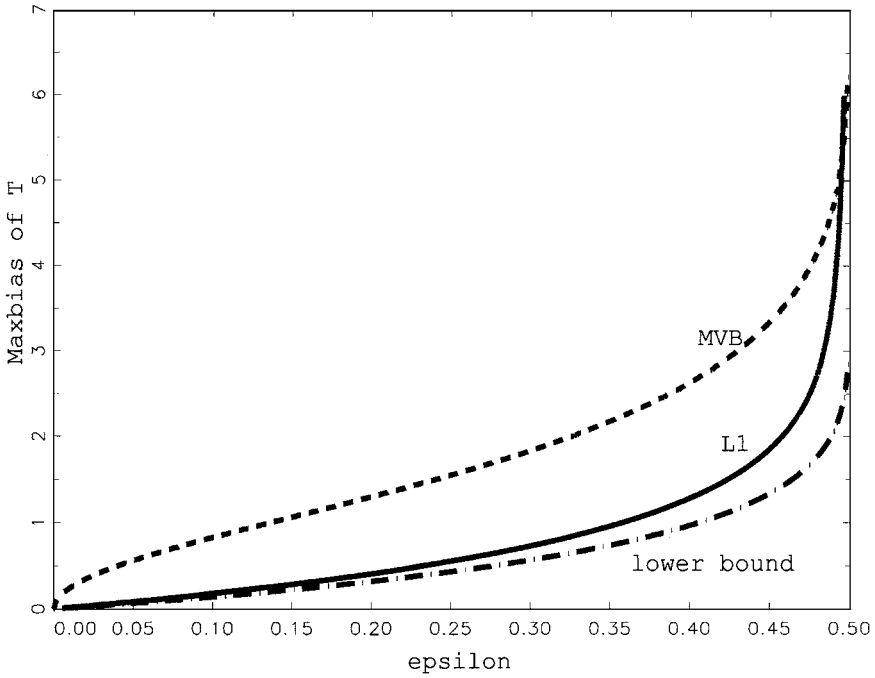

Fig. 1. *Maxbias curves of the location estimators L_1 and MVB together with the lower bound, for $p = 2$ and $\lambda_1 = \lambda_2 = 1$*

and Simpson (1993) *becomes*

$$\sqrt{\lambda_1}\Phi^{-1}\left(\frac{1}{2(1-\varepsilon)}\right). \tag{6.5}$$

In Fig. 1, we have plotted the maxbias curves of L_1 and MVB at $F = N(0, \mathbf{I}_2)$ together with the lower bound (6.5). We see that the L_1 estimator has a much lower maxbias curve than MVB. The maxbias curve of L_1 is very close to the lower bound and it is smooth in the neighborhood of $\varepsilon = 0$. On the other hand, the maxbias curve of MVB is not differentiable at $\varepsilon = 0$, and resembles the maxbias curve of LMS regression that was obtained by Martin, Yohai and Zamar (1989). Other values of $p \geq 2$ and $\lambda_1, \ldots, \lambda_p$ yield maxbias curves with comparable relative behaviors.

Note that this result seems to contradict the simulation results of the previous sections (Tables 1 and 3), where the MVB yielded a much smaller bias than L_1. This is because the maxbias curve is a "worst case" concept. In the simulation, the contaminating distribution was a point mass far away from the center of the model distribution. However, the "worst case" contaminating distribution for the MVB is a point mass much closer to this center, as can be seen from the proof of Theorem 2 part (*b*). Indeed, the MVB is a redescending estimator: observations far away from the bulk of the data have no influence on the MVB, whereas outliers somewhat closer to the center may affect the MVB.

7. A comparison with the Feasible Solution Algorithm

Exact computation of the MVE is possible (Cook, Hawkins and Weisberg 1993): consider all possible subsets of size h, called halfsamples, and compute the volume of the smallest ellipsoid covering all the points in the halfsample (which can be done in an exact manner by using the algorithm of Titterington 1975). The optimal halfsample has then the smallest value of all computed volumes, and the MVE is the ellipsoid associated with this optimal halfsample. Since the total number of all possible halfsamples is enormous, the exact computation is infeasible in practice. Unless the sample size is very small, one always needs to resort to approximative algorithms. In this paper, we focused on the $(p+1)$-subset algorithm, but more sophisticated algorithms exist, like heuristic search algorithms (Woodruff and Rocke 1993) or the Feasible Solution Algorithm (FSA, Hawkins 1993b). In this section we make a comparison with the FSA algorithm, which is well known and also used for computing the Least Median of Squares regression estimator (Hawkins 1993a).

The FSA starts from a randomly selected halfsample. Afterwards points in the halfsample are exchanged with points not belonging to it as long as this decreases the value of the objective function, i.e. the volume of the smallest ellipsoid covering all the points in the half-sample. If no further decrease is observed, then the obtained halfsample is called a "feasible solution". The algorithm considers a total number of *Nfsa* random starts. This number needs to be sufficiently high, certainly when applying FSA to noisy data sets. We used the implementation presently made public at the Statlib server (http://lib.stat.cmu.edu/general/) with $Nfsa = 50$ random starts (Hawkins 1993b used the same value of *Nfsa* for a data set with similar dimensions). The computation time required for the FSA is much higher than for the $(p+1)$-subset algorithm. A precise comparison of computation time is difficult, since tuning parameters need to be chosen by the user (*Nsamp*, *Nfsa*, h) and the

Table 4. *Using location adjustment for the FSA*

		Normal data			20% contaminated data		
		FSA	+ MVB	$+ L_1$	FSA	+ MVB	$+ L_1$
$p = 2$	Bias($\hat{\mu}$)	0.028	0.028	0.022	0.027	0.026	0.388
$n = 30$	MSE($\hat{\mu}$)	0.242	0.242	0.092	0.228	0.227	0.261
	$\text{med}_k \ \ln \phi_k$	0.606	0.606	0.606	0.385	0.385	0.385
	Ave_kObj_k	0.823	0.823	1.293	1.257	1.257	2.119
$p = 2$	Bias($\hat{\mu}$)	0.006	0.006	0.011	0.411	0.411	0.784
$n = 30$	MSE($\hat{\mu}$)	0.248	0.248	0.098	10.09	10.14	7.580
	$\text{med}_k \ \ln \phi_k$	0.973	0.973	0.973	0.682	0.682	0.682
	Ave_kObj_k	1.519	1.519	2.241	2.365	2.365	3.752

In each of four situations, simulating $m = 500$ samples yields the bias and MSE of the location estimators, the $\text{median}_{k=1,\ldots,m} \ln \phi_k$ of the scatter estimators, and the average value of the MVE objective function.

platforms on which the programs run may be different. For the values of n, *Nsamp* and *Nfsa* given in the simulation study, we may say that FSA is roughly 300 times slower than the $(p + 1)$-subset algorithm for $p = 2$ and 500 times for $p = 3$. This remains about the same when an L_1 adjustment is performed in the final step, since computing the L_1 estimator only once is very cheap. Adding the MVB location adjustment to the $(p + 1)$-subset algorithm makes the FSA about 150 times slower for $p = 2$, and 250 for $p = 3$. If the sample size and the dimension increase, we expect these differences in computation time to become even bigger. Note that we did not consider the case where the adjustment is made in every step, since this is computationally too expensive for the MVB, while giving not much extra gain (see discussion in Section 4).

While the computation time for the FSA is of a higher order of magnitude than for $(\hat{\mu}^p, \hat{\Sigma}^p)$ and for $(\tilde{\tilde{\mu}}, \tilde{\tilde{\Sigma}})$, the FSA succeeds in finding the lowest value of the objective function among the considered methods. This can be seen from the first column of Table 4, where the results of the simulation study of Section 4 are reported for the FSA. So we may say that the FSA attains its goal: achieving low values of the objective function (1.1). However the statistical benefits are rather limited: for the scatter part there is no significant difference with $(\hat{\mu}^p, \hat{\Sigma}^p)$ or with $(\tilde{\tilde{\mu}}, \tilde{\tilde{\Sigma}})$, and for the location part there is only a slight improvement in MSE. (Recall that standard errors around the simulated bias and MSE are about 0.01, and about 0.02 for the sphericity measures). Note that for $p = 3$ we have a huge bias and MSE under contamination. This is because breakdown occured in 8 out of the 500 runs. Increasing the value of *Nfsa* upto 100 did prevent this breakdown, but also doubled the computation time for the FSA.

Once the FSA solution of the MVE problem is obtained, one could think of improving the location estimate by adding an MVB or L_1 adjustment at the end. Since we will only need to do this once, it will not be very costly in comparison to the total computation time needed for the FSA. Table 4 reports the simulation results for FSA with location adjustment using MVB or L_1. Striking is that no difference can be observed between plain FSA and FSA + MVB. In fact, both procedures give very often (but not always) the same result. If the FSA has found the true MVE, then the MVB adjustment will of course not alter the estimate. Checking whether the MVB adjustment changed the location estimator can therefore be seen as a necessary, but not sufficient, condition to having found the exact MVE. Using an L_1 adjustment yields values comparable to $(\tilde{\tilde{\mu}}, \tilde{\tilde{\Sigma}})$ in Table 3: an increase of the value of the objective function, but a better statistical efficiency as measured by the MSE.

8. Conclusions

In this paper, we showed how the Minimum Volume Ball Estimator can be used for location adjustment of the Minimum Volume Ellipsoid Estimator. This adjustment always decreases the MVE objective function, but has little effect on the bias and MSE of $\hat{\mu}$. On the other hand, location adjustment based on L_1 does not necessarily lower the MVE objective function, but improves the efficiency of $\hat{\mu}$ for normal data.

In order to reduce the effect of extreme outliers on the L_1 adjustment, one can insert a reweighted L_1 estimator instead. Thus amounts to applying the L_1 estimator only to those observations which satisfy a certain condition. The algorithm given in Section 3 can easily be modified to compute this estimator. It suffices to replace step 2.4 by:

2.4" Compute the L_1 estimate $\theta_J := \mu_L(Y^*)$ where $Y^* = \{\mathbf{y}_i \in Y; (\mathbf{x}_i - \mu_J)^t S_J^{-1}(\mathbf{x}_i - \mu_J) \leq \frac{\chi^2_{p;0.975}}{\chi^2_{p;0.5}} \ \text{median}_j (\mathbf{x}_j - \mu_J)^t S_J^{-1}(\mathbf{x}_j - \mu_J)\}$ and calculate $R_J := \text{median}_i \|\mathbf{y}_i - \theta_J\|$.

As in the case of unweighted L_1, this type of adjustment does not lower the MVE objective function. According to some simulations, the bias of the resulting estimator is similar to that of the plain MVE estimator, while its MSE is better even in the contaminated situation.

An important advantage of the L_1 estimator is that its maxbias curve comes close to the lower bound, and is lower than the maxbias curve of the MVB location estimator. The situation

is completely analogous to the regression setup. We have seen in Section 4 that adjusting the intercept by the univariate LTS lowers the LTS objective function, but does not have much effect on the MSE of the coefficients. Adjusting the intercept by the univariate sample median yields a more efficient and low-bias intercept estimate.

Our recommendation is therefore to use the L_1 adjustment: it is cheap in computation time, has a low bias curve, and gives more efficient estimates of the multivariate location parameter in the normal case. Using MVB adjustment (or the Feasible Solution Algorithm which we discussed in Section 7) gives lower values of the objective function associated with the Minimum Volume Ellipsoid estimators, but at a high computational cost. Moreover, we also showed that lower values of the objective function do not guarantee a higher statistical precision.

Appendix

Proof of Theorem 1: Since

$$\min_{\mu} \operatorname*{median}_{i} \|x_i - \mu\| = \min_{\mu} \min_{1\le i_1<\cdots<i_h\le n} \max_{1\le j\le h} \|x_{i_j} - \mu\| = \min_{1\le i_1<\cdots<i_h\le n} \min_{\mu} \max_{1\le j\le h} \|x_{i_j} - \mu\| \tag{9.1}$$

the MVB is the smallest ball covering a certain subset of h points. Suppose w.l.o.g. that $i_j = j (j = 1, \ldots, h)$ yields the minimum in (9.1). Denote by $B(\hat{\mu}, \hat{R})$ the minimum volume ball covering $x_1, \ldots, x_h \in \mathbb{R}^p$. It is sufficient to prove that there exist k points $(2 \le k \le p + 1)$ such that

$$\begin{cases} \hat{\mu} \in L = \text{affinespan}\{x_1, \ldots, x_k\} & \text{with } \dim(L) = k - 1 \\ \|x_i - \hat{\mu}\| = \hat{R} & \text{for } i = 1, \ldots, k \end{cases} \tag{9.2}$$

The main tool used in this proof is the observation that all functions $\|x_i - \mu\| - R (i = 1, \ldots, h)$ are continuous in μ and R. Suppose that no point lies on the surface of the ball $B(\hat{\mu}, \hat{R})$. Then for any $\delta > 0$ sufficiently small, $B(\hat{\mu}, \hat{R} - \delta)$ still contains $x_1, \ldots, x_h$ but has a smaller volume, a contradiction. If one and only one point, say x_1, satisfies $\|x_1 - \hat{\mu}\| = \hat{R}$, consider the ball with center $\hat{\mu}^* = \hat{\mu} + \delta(x_1 - \hat{\mu})$ and radius $\hat{R}^* = \hat{R} - \delta\|x_1 - \hat{\mu}\| = (1 - \delta)\hat{R} < \hat{R}$. For $\delta > 0$ small enough, $B(\hat{\mu}^*, \hat{R}^*)$ will still contain $x_1, \ldots, x_h$, another contradiction. So, there exist at least two distinct points lying on the surface of the ball (remember that we assumed that no h observations can lie in a $(p - 1)$ dimensional hyperplane, hence at least two of these h observations must be different).

Let $m \ge 2$ be the number of observations $x_1, \ldots, x_m$ satisfying $\|x_i - \hat{\mu}\| = \hat{R}$, and select as many affinely independent points out of these as possible. Let us call them $x_1, \ldots, x_k$. Clearly, $k \ge 2$. If $k = p + 1$, then all the conditions stated in (9.2) are satisfied. On the other hand, if $k < p + 1$, let $L := \text{affinespan}\{x_1, \ldots, x_k\}$ hence $\dim(L) = k - 1$. The only condition which remains to be proved is that $\hat{\mu} \in L$. If $\hat{\mu} \notin L$, take $a = P_L(\hat{\mu}) - \hat{\mu}$ with $P_L(\hat{\mu})$ denoting the orthogonal projection of $\hat{\mu}$ onto L. Using orthogonality of a and $x_i - P_L(\hat{\mu})$ we obtain for any $\lambda > 0$ and for $1 \le i \le m$:

$$\begin{aligned} \|x_i - (\hat{\mu} + \lambda a)\|^2 &= \|x_i - P_L(\hat{\mu})\|^2 + \|P_L(\hat{\mu}) - (\hat{\mu} + \lambda a)\|^2 \\ &= \|x_i - P_L(\hat{\mu})\|^2 + (1 - \lambda)^2\|a\|^2 \\ &= \hat{R}^2 + \lambda(\lambda - 2)\|a\|^2 \end{aligned}$$

For λ sufficiently small, the ball $B(\hat{\mu}^*, \hat{R}^*)$ where $\hat{\mu}^* = \hat{\mu} + \lambda a$ and $\hat{R}^{*2} = \hat{R}^2 + \lambda(\lambda - 2)\|a\|^2$ has a smaller radius than $B(\hat{\mu}, \hat{R})$ while still containing all the points, a contradiction. Hence, $\hat{\mu} \in L$. □

Proof of Theorem 2, part (a): If G denotes a distribution on $\mathbb{R}^p$, the L_1 estimator is given by the functional $T(G)$ satisfying

$$T(G) = \operatorname*{argmin}_{t} \int (\|x - t\| - \|x\|)\, dG(x). \tag{9.3}$$

Therefore, $T(G)$ satisfies $\int u(x - t)\, dG(x) = 0$ where $u(y) = \frac{y}{\|y\|}$ if $y \ne 0$ and 0 otherwise. Let $F = N(\mathbf{0}, \Sigma)$, with $\Sigma = \text{diag}(\lambda_1, \ldots, \lambda_p)$, $\lambda_1 \ge \cdots \ge \lambda_p > 0$. Under contamination, the model becomes $F_\varepsilon = (1 - \varepsilon)F + \varepsilon H$ where H can be any distribution and the functional $T(F_\varepsilon)$ satisfies $\int u(x - t)\, dF_\varepsilon(x) = 0 \Leftrightarrow (1-\varepsilon)\int u(x-t)\, dF(x) + \varepsilon \int u(x-t)\, dH(x) = 0$. By definition of L_1 and due to the form of F, the most influential distribution H is given by a point mass placed at infinity on the first axis. Replacing H by $\Delta_{\xi e_1}$, where e_1 is the first unit vector, yields

$$(1 - \varepsilon) \int u(x - t)\, dF(x) + \varepsilon u(\xi e_1 - t) = 0. \tag{9.4}$$

Letting ξ tend to $+\infty$ shows that $T(F_\varepsilon) = T((1-\varepsilon)F + \varepsilon\Delta_{\infty e_1})$ satisfies

$$(1 - \varepsilon) \int u(x - T(F_\varepsilon))\, dF(x) + \varepsilon e_1 = 0$$

$$\Rightarrow \begin{cases} \int u(x - T(F_\varepsilon))_1\, dF(x) = -\frac{\varepsilon}{1-\varepsilon} & \text{(i)} \\ \int u(x - T(F_\varepsilon))_j\, dF(x) = 0, \quad \text{for } 2 \le j \le p & \text{(ii)} \end{cases}$$

Due to the symmetry of F, only the first coordinate of $T(F_\varepsilon)$ differs from zero and equation (ii) is trivially satisfied when $T(F_\varepsilon)_j = 0$. On the other hand, equation (i) yields

$$\int \frac{(x_1 - T(F_\varepsilon)_1)}{\|x - T(F_\varepsilon)_1 e_1\|}\, dF(x) = -\frac{\varepsilon}{1 - \varepsilon} \tag{9.5}$$

Denoting $T(F_\varepsilon)_1 = b$ and φ the probability density function of the standard normal distribution, the right hand side of (9.5) becomes:

$$\begin{aligned} &\int \cdots \int \frac{x_1 - b}{\sqrt{(x_1 - b)^2 + x_2^2 + \cdots + x_p^2}} f(x_1) \ldots f(x_p)\, dx_1 \ldots dx_p \\ &= \frac{1}{\sqrt{\lambda_1 \ldots \lambda_p}} \int \cdots \int \frac{x_1 - b}{\sqrt{(x_1 - b)^2 + x_2^2 + \cdots + x_p^2}} \\ &\quad \times \varphi\left(\frac{x_1}{\sqrt{\lambda_1}}\right) \ldots \varphi\left(\frac{x_p}{\sqrt{\lambda_p}}\right) dx_1 \ldots dx_p \end{aligned}$$

$$= \int \cdots \int \frac{y_1\sqrt{\lambda_1} - b}{\sqrt{(y_1\sqrt{\lambda_1} - b)^2 + \lambda_2 y_2^2 + \cdots + \lambda_p y_p^2}} \times \varphi(y_1)\ldots\varphi(y_p)\,dy_1\ldots dy_p$$

where $Y_1, \ldots, Y_p \overset{\text{iid}}{\sim} N(0, 1)$. Letting $V = \sum_{i=2}^{p} \lambda_i Y_i^2$ and $Z = Y_1\sqrt{\lambda_1} - b$, equation (i) becomes:

$$\frac{1}{\sqrt{\lambda_1}} \int_{-\infty}^{+\infty} \int_{0}^{+\infty} \frac{z}{\sqrt{z^2 + v}} \varphi\left(\frac{z+b}{\sqrt{\lambda_1}}\right) g_V(v)\,dz\,dv = -\frac{\varepsilon}{1-\varepsilon}$$

where g_V is the probability density function of V. □

Proof of Theorem 2, part (b): If G denotes an arbitrary distribution on $\mathbb{R}^p$, the Minimum Volume Ball Estimator is given by the functional $T(G)$ which is the first argument of (μ, R) minimizing R subject to $\mathrm{P}_G(X \in B(\mu, R)) \geq \frac{1}{2}$. Here, $B(\mu, R)$ represents the ball with center μ and radius R. Let $F = N(\mathbf{0}, \Sigma)$, with $\Sigma = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$, $\lambda_1 \geq \cdots \geq \lambda_p > 0$. We have $T(F) = \mathbf{0}$ and

$$\mathrm{P}_F(X \in B(\mathbf{0}, R)) = \int_{\|x\| \leq R} dF(x) = \frac{1}{\sqrt{\det(\Sigma)}} \int_{\|x\| \leq R} g(x^t \Sigma^{-1} x)\,dx,$$

where $g(t) = (\frac{1}{\sqrt{2\pi}})^p e^{-\frac{t}{2}}$.

Since the bias will be largest if we contaminate at a point in the direction of the first unit vector e_1, we can take F_ε of the form $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_{\xi e_1}$, with w.l.o.g. $\xi > 0$. The maxbias curve of the MVB estimator is now given by

$$B(\varepsilon, \mathrm{MVB}, F) = \sup_{\xi} \|T(F_\varepsilon)\| = \sup_{\xi} |T(F_\varepsilon)_1| \tag{9.6}$$

since $T(F_\varepsilon)_k = 0, \forall k > 1$ due to symmetry. For ξ fixed, the MVB is given by $(\mu_\varepsilon(\xi)e_1, R_\varepsilon(\xi))$ minimizing the second argument among all $(\mu e_1, R)$ satisfying

$$\mathrm{P}_{F_\varepsilon}(X \in B(\mu e_1, R)) \geq \frac{1}{2} \Leftrightarrow (1-\varepsilon)\frac{1}{\sqrt{\det(\Sigma)}} \times \int_{\|x - \mu e_1\| \leq R} g(x^t \Sigma^{-1} x)\,dx + \varepsilon I(|\xi - \mu| \leq R) \geq \frac{1}{2}. \tag{9.7}$$

We can limit ourselves to three possible solutions for the above problem:

- type I: $\mu_\varepsilon(\xi) = 0$ and $R_\varepsilon(\xi) < \xi$ (i.e. the point contamination is outside the ball $B(\mu_\varepsilon(\xi)e_1, R_\varepsilon(\xi))$).
- type II: $\mu_\varepsilon(\xi) = 0$ and $R_\varepsilon(\xi) \geq \xi$ (i.e. the point contamination is inside the ball $B(\mu_\varepsilon(\xi)e_1, R_\varepsilon(\xi))$).
- type III: $\mu_\varepsilon(\xi) > 0$ and $R_\varepsilon(\xi) = \xi - \mu_\varepsilon(\xi)$ (i.e. the point contamination is on the edge of the ball $B(\mu_\varepsilon(\xi)e_1, R_\varepsilon(\xi))$). This case assumes that $R_\varepsilon(\xi) \leq \xi$ since $\mu_\varepsilon(\xi) > 0$. The other case, $\mu_\varepsilon(\xi) \leq 0$, is not worth investigating since type II attains in this case a lower value for the objective function.

In the first situation, the constraint (9.7) becomes

$$(1-\varepsilon)\frac{1}{\sqrt{\det(\Sigma)}} \int_{\|x\| \leq R} g(x^t \Sigma^{-1} x)\,dx \geq \frac{1}{2}$$

and $R_\varepsilon(\xi)$ would be given by the solution of

$$\int_{\|x\| \leq R} g(x^t \Sigma^{-1} x)\,dx = \frac{\sqrt{\det(\Sigma)}}{2(1-\varepsilon)}, \tag{9.8}$$

which is independent from ξ and will be denoted by R_ε^+. Similarly, type II would yield an $R_\varepsilon(\xi) = R_\varepsilon^-$ defined by

$$\int_{\|x\| \leq R} g(x^t \Sigma^{-1} x)\,dx = \frac{\sqrt{\det(\Sigma)}(1-2\varepsilon)}{2(1-\varepsilon)}. \tag{9.9}$$

For type III, we only need to consider values of ξ greater than R_ε^-. Replacing $\mu_\varepsilon(\xi)$ by $\xi - R_\varepsilon(\xi)$ in the constraint shows that $R_\varepsilon(\xi)$ would be a solution of the equation

$$\int_{\|x - (\xi - R)e_1\| \leq R} g(x^t \Sigma^{-1} x)\,dx = \frac{\sqrt{\det(\Sigma)}(1-2\varepsilon)}{2(1-\varepsilon)}. \tag{9.10}$$

For any $\xi \geq R_\varepsilon^-$, such a solution will be denoted by $\tilde{R}_\varepsilon(\xi)$.

One can easily verify that $R_\varepsilon^- < R_\varepsilon^+$, $\tilde{R}_\varepsilon(R_\varepsilon^-) = R_\varepsilon^-$, $\tilde{R}_\varepsilon(\xi) \leq \xi$ and that $\tilde{R}_\varepsilon(\xi)$ increases strictly in ξ. Combining all these results, we obtain:

$$\begin{cases} (\mu_\varepsilon(\xi)e_1, R_\varepsilon(\xi)) = (0, R_\varepsilon^-) & \text{for } \xi < R_\varepsilon^- \\ (\mu_\varepsilon(\xi)e_1, R_\varepsilon(\xi)) = ((\xi - \tilde{R}_\varepsilon(\xi))e_1, \tilde{R}_\varepsilon(\xi)) & \text{for } R_\varepsilon^- \leq \xi < R_\varepsilon^+ \\ (\mu_\varepsilon(\xi)e_1, R_\varepsilon(\xi)) = \begin{cases} ((\xi - \tilde{R}_\varepsilon(\xi))e_1, \tilde{R}_\varepsilon(\xi)) & \text{if } \tilde{R}_\varepsilon(\xi) \leq R_\varepsilon^+ \\ (0, R_\varepsilon^+) & \text{otherwise} \end{cases} & \text{for } \xi \geq R_\varepsilon^+ \end{cases} \tag{9.11}$$

The bias is either equal to 0 (type I and II) or to $\xi - \tilde{R}_\varepsilon(\xi)$ (type III) which strictly increases with respect to ξ. The maximal bias (9.6) equals thus $\xi^* - \tilde{R}_\varepsilon(\xi^*)$ where ξ^* satisfies

$$\tilde{R}_\varepsilon(\xi^*) = R_\varepsilon^+ \Leftrightarrow \int_{\|x - (\xi^* - R_\varepsilon^+)e_1\| \leq R_\varepsilon^+} g(x^t \Sigma^{-1} x)\,dx = \frac{\sqrt{\det(\Sigma)}(1-2\varepsilon)}{2(1-\varepsilon)}.$$

Therefore, the maximal bias $B(\varepsilon, \mathrm{MVB}, F)$ is the solution b of

$$\int_{\|x - be_1\| \leq R_\varepsilon^+} g(x^t \Sigma^{-1} x)\,dx = \frac{\sqrt{\det(\Sigma)}(1-2\varepsilon)}{2(1-\varepsilon)}. \tag{9.12}$$

Transforming the variable x to $y = x - be_1$, equation (6.3) follows from (9.12). □

Proof of Theorem 2, part (c): According to Theorem 2.1 in He and Simpson (1993), a lower bound for the maxbias $B(\varepsilon, T, F)$ at the normal model $\{F_\theta = N(\theta, \mathrm{diag}(\lambda_1, \ldots, \lambda_p)) \mid \theta \in \mathbb{R}^p\}$ is given by the solution b_0 of

$$\inf_{\|\mathbf{z}\| = 1} d_v\big(F_0, F_{2b_0\mathbf{z}}\big) = \frac{\varepsilon}{1-\varepsilon} \tag{9.13}$$

where the variation distance $d_v(F_0, F_{2b_0\mathbf{z}})$ equals $\int(f(\mathbf{x}) - f(\mathbf{x}-2b_0\mathbf{z}))_+ d\mathbf{x}$. Denote $H_\mathbf{z} = \{\mathbf{x} \in \mathbb{R}^p \mid f(\mathbf{x}) \geq f(\mathbf{x}-2b_0\mathbf{z})\} = \{\mathbf{x} \in \mathbb{R}^p \mid \mathbf{z}^t\Sigma^{-1}\mathbf{x} \leq b_0\mathbf{z}^t\Sigma^{-1}\mathbf{z}\}$. Then,

$$\begin{aligned} d_v\big(F_0, F_{2b_0\mathbf{z}}\big) &= P_F(H_\mathbf{z}) - P_F(H_\mathbf{z} - 2b_0\mathbf{z}) \\ &= P_F(\{\mathbf{x} \in \mathbb{R}^p \mid -b_0\mathbf{z}^t\Sigma^{-1}\mathbf{z} \leq \mathbf{z}^t\Sigma^{-1}\mathbf{x} \\ &\leq b_0\mathbf{z}^t\Sigma^{-1}\mathbf{z}\}) \\ &= P_0(\{\mathbf{x} \in \mathbb{R}^p \mid |(\Sigma^{-\frac{1}{2}}\mathbf{z})^t\mathbf{x}| \leq b_0\mathbf{z}^t\Sigma^{-1}\mathbf{z}\}), \end{aligned}$$

where $P_0 = N(\mathbf{0}, \mathbf{I})$.

Due to $\|\mathbf{z}\| = 1$ and the symmetry of P_0,

$$\begin{aligned} d_v\big(F_0, F_{2b_0\mathbf{z}}\big) &= \Phi(\{y \in \mathbb{R} \mid |y| \leq b_0\sqrt{\mathbf{z}^t\Sigma^{-1}\mathbf{z}}\}) \\ &= 2\Phi(b_0\sqrt{\mathbf{z}^t\Sigma^{-1}\mathbf{z}}) - 1 \end{aligned} \tag{9.14}$$

which becomes minimal for $\mathbf{z}$ equal to the normalized eigenvector corresponding to the smallest eigenvalue of Σ^{-1}, yielding

$$\inf_{\|\mathbf{z}\|=1} d_v\big(F_0, F_{2b_0\mathbf{z}}\big) = 2\Phi\left(\frac{b_0}{\sqrt{\lambda_1}}\right) - 1. \tag{9.15}$$

Combining (9.13) and (9.15) yields

$$b_0 = \sqrt{\lambda_1}\,\Phi^{-1}\left(\frac{1}{2(1-\varepsilon)}\right). \tag{9.16}$$

□

References

Cook R.D., Hawkins D.M., and Weisberg S. 1993. Exact iterative computation of the robust multivariate minimum volume ellipsoid estimator. Statistics and Probability Letters 16: 213–218.

Hawkins D.M. 1993a. The Feasible Set Algorithm for least median of squares regression. Computational Statistics and Data Analysis 16: 81–101.

Hawkins D.M. 1993b. A Feasible Solution Algorithm for the minimum volume ellipsoid estimator in multivariate data. Computational Statistics 8: 95–107.

He X. and Simpson D.G. 1993. Lower bounds for contamination bias: Globally minimax versus locally linear estimation. The Annals of Statistics 21: 314–337.

Hössjer O. and Croux C. 1995. Generalizing univariate signed rank statistics for testing and estimating a multivariate location parameter. Nonparametric Statistics 4: 293–308.

Lopuhaä H.P. 1992. Highly efficient estimators of multivariate location with high breakdown point. The Annals of Statistics 20: 398–413.

Lopuhaä H.P. and Rousseeuw P.J. 1991. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. The Annals of Statistics 19: 229–248.

Maronna R.A. and Yohai V.J. 1995. The behavior of the Stahel-Donoho robust multivariate estimator. Journal of the American Statistical Association 90: 330–341.

Martin R.D., Yohai V.J., and Zamar R.H. 1989. Min-max bias robust regression. The Annals of Statistics 17: 1608–1630.

Rousseeuw P.J. 1984. Least median of squares regression. Journal of the American Statistical Association 79: 871–880.

Rousseeuw P.J. 1985. Multivariate estimation with high breakdown point. In: Grossmann W., Pflug G., Vincze I., and Wertz W. (Eds.), Mathematical Statistics and Applications, Vol. B, Reidel, Dordrecht, pp. 283–297.

Rousseeuw P.J. and Leroy A.M. 1987. Robust Regression and Outlier Detection. John Wiley, New York.

Rousseeuw P.J. and van Zomeren B.C. 1990. Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association 85: 633–639.

Titterington D.M. 1975. Optimal design: Some geometrical aspects of D-optimality. Biometrika 62: 313–320.

Woodruff D.L. and Rocke D.M. 1993. Heuristic search algorithms for the minimum volume ellipsoid. Journal of Computational and Graphical Statistics 2: 69–95.