KATHOLIEKE
UNIVERSITEIT
LEUVEN

DEPARTEMENT TOEGEPASTE ECONOMISCHE WETENSCHAPPEN

RESEARCH REPORT 0224

THE MULTIVARIATE LEAST TRIMMED SQUARES ESTIMATOR

by
J. AGULLO
C. CROUX
S. VAN AELST

D/2002/2376/24

The Multivariate Least Trimmed Squares Estimator

Jose Agulló[1], Christophe Croux,[2] and Stefan Van Aelst,[3]

Abstract

In this paper we introduce the least trimmed squares estimator for multivariate regression. We give three equivalent formulations of the estimator and obtain its breakdown point. A fast algorithm for its computation is proposed. We prove Fisher-consistency at the multivariate regression model with elliptically symmetric error distribution and derive the influence function. Simulations investigate the finite-sample efficiency and robustness of the estimator. To increase the efficiency of the estimator, we also consider a one-step reweighted version, as well as multivariate generalizations of one-step GM-estimators.

Keywords: Multivariate Regression, Breakdown Point, Generalized M-estimator, Influence Function, Minimum Covariance Determinant Estimator.

1 Introduction

Consider the multivariate regression model

$$y_i = \mathcal{B}^t x_i + \varepsilon_i$$

$i = 1, \ldots, n$ with $x_i = (x_{i1}, \ldots, x_{ip})^t \in I\!R^p$ and $y_i = (y_{i1}, \ldots, y_{iq})^t \in I\!R^q$. The matrix $\mathcal{B} \in I\!R^{p \times q}$ contains the regression coefficients. The error terms $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. with zero center and as scatter a positive definite and symmetric matrix Σ of size q. Furthermore, we assume that the errors are independent of the carriers. Note that this model generalizes both the univariate regression model ($q = 1$) and the multivariate location model ($x_i = 1$). Denote the entire sample $Z_n = \{(x_i, y_i); i = 1, \ldots, n\}$ and write $X = (x_1, \ldots, x_n)^t$ for the

[1] Departamento de Fundamentos del Análisis Económico, University of Alicante, E-03080, Alicante, Spain

[2] Dept. of Applied Economics, KULeuven, Naamsestraat 69, B-3000 Leuven, Belgium.

[3] Dept. of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 S9, B-9000 Gent, Belgium.

design matrix and $Y = (y_1, \ldots, y_n)^t$ for the matrix of responses. The classical estimator for $\mathcal{B}$ is the least-squares (LS) estimator $\mathcal{B}_{LS}$ which is given by

$$\hat{\mathcal{B}}_{LS} = (X^t X)^{-1} X^t Y \tag{1.1}$$

while Σ is unbiasedly estimated by

$$\hat{\Sigma}_{LS} = \frac{1}{n-p}(Y - X\hat{\mathcal{B}}_{LS})^t (Y - X\hat{\mathcal{B}}_{LS}). \tag{1.2}$$

Since the least squares estimator is extremely sensitive to outliers, we aim to construct a robust alternative. An overview of strategies to robustify the multivariate regression method is given by Maronna and Yohai (1997) in the context of simultaneous equations models. Koenker and Portnoy (1990) apply a regression M-estimator to each coordinate of the responses and Bai et al. (1990) minimize the sum of the euclidean norm of the residuals. However, these two methods are not affine equivariant. Methods based on robust estimation of the location and scatter of the joint distribution of the (x, y) variables have been introduced by Ollila et al. (2001,2002) who use rank and sign based covariance matrices and by Rousseeuw et al. (2000) who use the Minimum Covariance Determinant estimator (Rousseeuw 1984). Our approach will be different from the latter, since it will be based on the covariance matrix of the residuals, more than on the covariance matrix of the joint distribution.

In Section 2 we give a formal definition of the multivariate least trimmed squares (MLTS) estimator and derive two equivalent formulations allowing us to study more easily the properties of the estimator. In Section 3 we show that the estimator has a positive breakdown point (BDP). A time efficient algorithm to compute the MLTS is presented in Section 4. In Section 5 we give a functional version of the multivariate least trimmed squares estimator and show that the estimator is Fisher-consistent at the multivariate regression model with elliptically symmetric error distribution. Afterwards, in section 6 we derive its influence function and compute asymptotic variances and corresponding efficiencies. In section 7 we consider a one-step reweighted version of the estimator as well as a multivariate generalization of one-step GM estimators with Mallows and Schweppe type weights using the MLTS as initial estimator. Section 8 presents simulation results. Simulations have been done to investigate the finite-sample efficiency and robustness of the MLTS estimator. Section 9 presents a real data example while Section 10 concludes. The Appendix contains all the proofs.

2 Definition and properties

Our approach consists of finding the subset of h observations having the property that the determinant of the covariance matrix of its residuals from a LS-fit solely based on this subset is minimal. The resulting estimator will then be simply the LS-estimator computed from the optimal subset. The definition of the estimator is reminiscent of that of the MCD location/scatter estimator of Rousseeuw (1984), and reduces to it in case of a multivariate regression model with only an intercept, where $X = (1, \ldots, 1)^t \in I\!R^n$. Indeed in the latter case the multivariate regression model reduces to a multivariate location model. We will show that our approach is equivalent to the selection of the value of $\mathcal{B}$ which minimizes the determinant of the robust MCD scatter matrix of the residuals. Of course, one could also think of minimizing the determinant of other robust covariance matrices of the residuals. As such, Bilodeau and Duchesne (2000) used S-estimators as robust estimator of the covariance of the residuals in the context of seemingly unrelated regression. We thus use the Minimum Covariance Determinant estimator (MCD) as scatter matrix estimator of the residuals. The main reason for this choice is that it turns out to be easy to develop a fast algorithm for the resulting multivariate regression estimator. Moreover, the resulting estimator has a high BDP and is ideally suited as initial estimator for one (or more) step procedures.

Consider a dataset $Z_n = \{(x_i, y_i); i = 1, \ldots, n\} \subset I\!R^{p+q}$ and for any $\mathcal{B} \in I\!R^{p \times q}$ denote $r_i(\mathcal{B}) = y_i - \mathcal{B}^t x_i$ the corresponding residuals. Let $\mathcal{H} = \{H \subset \{1, \ldots, n\} | \#H = h\}$ be the collection of all subsets of size h. For any $H \in \mathcal{H}$ denote $\hat{\mathcal{B}}_{LS}(H)$ the least squares fit based solely on the observations $\{(x_j, y_j); j \in H\}$. Furthermore, for any $H \in \mathcal{H}$ and $\mathcal{B} \in I\!R^{p \times q}$ denote $\text{cov}(H, \mathcal{B}) := \frac{1}{h} \sum_{j \in H} (r_j(\mathcal{B}) - \bar{r}_H(\mathcal{B}))(r_j(\mathcal{B}) - \bar{r}_H(\mathcal{B}))^t$, with $\bar{r}_H(\mathcal{B}) := \frac{1}{h} \sum_{j \in H} r_j(\mathcal{B})$, the covariance matrix of the residuals with respect to the fit $\mathcal{B}$, belonging to the subset H. Then the MLTS estimator is defined as follows:

Definition 1. *With the notations above the multivariate least trimmed squares estimator (MLTS) is defined as*

$$\hat{\mathcal{B}}_{MLTS}(Z_n) = \hat{\mathcal{B}}_{LS}(\hat{H}) \text{ where } \hat{H} \in \underset{H \in \mathcal{H}}{\operatorname{argmin}} \det \hat{\Sigma}_{LS}(H) \tag{2.1}$$

with $\hat{\Sigma}_{LS}(H) = \text{cov}(H, \hat{\mathcal{B}}_{LS}(H))$ *for any* $H \in \mathcal{H}$. *The covariance of the errors can then be estimated by*

$$\hat{\Sigma}_{MLTS}(Z_n) = c_\alpha \hat{\Sigma}_{LS}(\hat{H}), \tag{2.2}$$

where c_α *is a consistency factor.*

Note that if the minimization problem has more than one solution, in which case we look at $\text{argmin}_H \det \hat{\Sigma}_{LS}(H)$ as a set, we arbitrarily select one of these solutions to determine the MLTS estimator. In Section 5 a consistency factor c_α will be proposed to attain Fisher-consistency at the specified model. Note that for $h = n$ we find back the classical least squares estimator. Throughout the text we will suppose that the dataset $Z_n = \{(x_i, y_i); i = 1, \ldots, n\} \subset I\!R^{p+q}$ is in *general position* in the sense that no h points of Z_n are lying on the same hyperplane of $I\!R^{p+q}$. Formally, this means that for all $\beta \in I\!R^p$, $\gamma \in I\!R^q$, it holds that

$$\#\{(x_j, y_j) \,|\, \beta^t x_j + \gamma^t y_j = 0\} < h \tag{2.3}$$

unless if β and γ are both zero vectors. For datasets in general position we will now give two equivalent characterizations of the MLTS estimator. First, we need the following lemma which is a generalization of the characterization of Grübel (1988) of the mean and covariance matrix of a multivariate distribution.

Lemma 1. *Let $z = (x, y)$ be a $(p+q)$-dimensional random variable having distribution K. Suppose that $E_K[xx^t]$ is a strictly positive definite matrix. Define $\mathcal{B}_{LS}(K) = E_K[xx^t]^{-1}E_K[xy^t]$ and $\Sigma_{LS}(K) = \text{Cov}_0(\varepsilon) := E_K[\varepsilon\varepsilon^t]$ where $\varepsilon := y - (\mathcal{B}_{LS}(K))^t x$. Then among all pairs (b, Δ) with $b \in I\!R^{p\times q}$ and Δ a positive definite and symmetric matrix of size q such that*

$$E_K[(y - b^t x)^t \Delta^{-1} (y - b^t x)] = q, \tag{2.4}$$

the unique pair which minimizes $\det \Delta$ is given by $(\mathcal{B}_{LS}(K), \Sigma_{LS}(K))$.

Note that if not all points of a dataset are lying in a subspace of $I\!R^{p+q}$, then Lemma 1 can be applied by taking for K the empirical distribution function associated to the data. This results in a characterization of the *sample* least squares estimators for the multivariate regression model.

Now we are ready to show that the MLTS estimator can also be obtained as the $\mathcal{B}$ minimizing the determinant of the MCD scatter matrix estimate computed from its residuals. Herefore, denote $\text{MCD}_q(\mathcal{B})$ the MCD-scatter matrix based on the residuals from $\mathcal{B}$. Formally,

$$\text{MCD}_q(\mathcal{B}) = \text{Cov}_0(\hat{H}, \mathcal{B}) = \frac{1}{h} \sum_{j \in \hat{H}} r_j(\mathcal{B}) r_j(\mathcal{B})^t$$

with $\hat{H} \in \underset{H \in \mathcal{H}}{\text{argmin}} \det \text{Cov}_0(H, \mathcal{B})$ for any $H \in \mathcal{H}$ and $\mathcal{B} \in I\!R^{p\times q}$. The residual covariance matrices we consider are thus centered at zero. (If we work with a model with intercept it can be shown that "Cov_0" may be replaced by the usual sample covariance matrix of the residuals.)

Proposition 1. *With the notations above, for datasets in general position, we have that*

$$\underset{\mathcal{B}}{\operatorname{argmin}} \det \mathrm{MCD}_q(\mathcal{B}) = \{\hat{\mathcal{B}}_{LS}(\hat{H}) \;|\hat{H} \in \underset{H\in\mathcal{H}}{\operatorname{argmin}} \det \hat{\Sigma}_{LS}(H)\} \tag{2.5}$$

Proposition 1 shows that any $\mathcal{B}$ which minimizes the determinant of the MCD scatter estimate of its residuals is also a solution of (2.1). In the case of unique solutions, which we have almost surely if we sample from a continuous distribution, we can rewrite (2.5) as

$$\hat{\mathcal{B}}_{MLTS}(Z_n) = \underset{\mathcal{B}}{\operatorname{argmin}} \det \mathrm{MCD}_q(\mathcal{B}). \tag{2.6}$$

For the residual scatter estimator we have

$$\hat{\Sigma}_{MLTS}(Z_n) = c_\alpha \mathrm{MCD}_q(\hat{\mathcal{B}}_{MLTS}(Z_n)) \tag{2.7}$$

A third characterization of the MLTS is based on the distances of the residuals. For any $\mathcal{B} \in I\!R^{p\times q}$ and $\Sigma \in \mathrm{PDS}(q)$, the set of positive definite and symmetric matrices of size q, we define the squared distances (for the Σ metric) of the residuals w.r.t. $\mathcal{B}$ as

$$d_i^2(\mathcal{B},\Sigma) := r_i(\mathcal{B})^t\Sigma^{-1}r_i(\mathcal{B}).$$

Denote $d_{1:n}(\mathcal{B},\Sigma) \leq \cdots \leq d_{n:n}(\mathcal{B},\Sigma)$ the ordered sequence of distances of the residuals. Then the MLTS estimator can also be obtained in the following way.

Proposition 2. *Consider*

$$\underset{\mathcal{B},\Sigma;|\Sigma|=1}{\operatorname{argmin}} \sum_{j=1}^{h} d_{j:n}^2(\mathcal{B},\Sigma) \tag{2.8}$$

where the minimum is over all $\mathcal{B} \in I\!R^{p\times q}$ and $\Sigma \in \mathrm{PDS}(q)$ with $\det\Sigma = 1$ (denoted as $|\Sigma| = 1$). Then for datasets in general position it holds that

$$\left\{\tilde{\mathcal{B}} \,|(\tilde{\mathcal{B}},\tilde{\Sigma}) \in \underset{\mathcal{B},\Sigma;|\Sigma|=1}{\operatorname{argmin}} \sum_{j=1}^{h} d_{j:n}^2(\mathcal{B},\Sigma)\right\} = \{\hat{\mathcal{B}}_{LS}(\hat{H}) \;|\hat{H} \in \underset{H}{\operatorname{argmin}} \det \hat{\Sigma}_{LS}(H)\} \tag{2.9}$$

Proposition 2 shows that any $\tilde{\mathcal{B}}$ minimizing the sum of the h smallest squared distances of its residuals (subject to $\det\Sigma = 1$) is also a solution of (2.1). For any $(\tilde{\mathcal{B}},\tilde{\Sigma})$ that minimizes (2.8) denote $\tilde{H} := \{j; d_j^2(\tilde{\mathcal{B}},\tilde{\Sigma}) \leq d_{h:n}^2(\tilde{\mathcal{B}},\tilde{\Sigma})\} \in \mathcal{H}$ the set of indices corresponding to the h smallest squared distances of the residuals. In the case of unique solutions, Proposition 2 yields

$$\hat{\mathcal{B}}_{MLTS}(Z_n) = \underset{\mathcal{B},\Sigma;|\Sigma|=1}{\operatorname{argmin}} \sum_{j=1}^{h} d_{j:n}^2(\mathcal{B},\Sigma), \tag{2.10}$$

so the MLTS estimator minimizes the sum of the h smallest squared distances of its residuals (subject to the condition $\det \Sigma = 1$). Note that in the case $q = 1$ expression (2.8) reduces to $\text{argmin}_{\mathcal{B}} \sum_{j=1}^{h} r_{j:n}^2(\mathcal{B})$, with $r_{1:n}(\mathcal{B}) \leq \cdots \leq r_{n:n}(\mathcal{B})$ the ordered residuals w.r.t. $\mathcal{B}$. Hence in the case of univariate regression our estimator minimizes the sum of the h smallest squared residuals, and thus corresponds to the Least Trimmed Squares estimator (LTS) of (Rousseeuw 1984). This explains why we call our estimator the MLTS estimator. The LTS is a well-known positive-breakdown robust estimator for regression which is frequently used.

3 Breakdown point

To study the global robustness of the MLTS estimator we compute its finite-sample breakdown point. The finite-sample breakdown point ε_n^* of an estimator T_n is the smallest fraction of observations from Z_n that need to be replaced by arbitrary values to carry the estimate beyond all bounds (Donoho and Huber 1983). Formally, it is defined as

$$\varepsilon_n^*(T_n, Z_n) = \min\{\frac{m}{n}; \sup_{Z_n'} \|T_n(Z_n) - T_n(Z_n')\| = \infty\}$$

where the supremum is over all possible collections Z_n' obtained from Z_n by replacing m data points by arbitrary values. For any dataset $Z_n \subset I\!R^{p+q}$ denote $k(Z_n)$ the maximal number of observations of Z_n lying on a same hyperplane of $I\!R^{p+q}$. Since we required that Z_n is in general position, we have $k(Z_n) < h$. We now have the following theorem.

Theorem 1. *For any dataset $Z_n \subset I\!R^{p+q}$ in general position with $q > 1$ it holds that*

$$\varepsilon_n^*(\hat{\mathcal{B}}_{MLTS}, Z_n) = \frac{\min(n - h + 1, h - k(Z_n))}{n}. \tag{3.1}$$

It follows that if we take $h = \gamma n$ for some fraction $0 < \gamma \leq 1$ then the corresponding breakdown point equals $\varepsilon_n^*(\hat{\mathcal{B}}_{MLTS}, Z_n) = \min(1 - \gamma + 1/n, \gamma - k(Z_n)/n)$. If the dataset Z_n comes from a continuous distribution F, then with probability 1, no $p + q$ points belong to the same hyperplane of $I\!R^{p+q}$. This implies $k(Z_n) = p + q - 1$ and $\varepsilon_n^*(\hat{\mathcal{B}}_{MLTS}, Z_n) = \min(n - h + 1, h - p - q + 1)/n$ almost surely. Then for $h = \gamma n$ the breakdown point of the MLTS tends to $\min(1 - \gamma, \gamma)$. It follows that for data with $k(Z_n) = p + q - 1$ any choice $[(n + p + q)/2] \leq h \leq [(n + p + q + 1)/2]$ yields the maximal breakdown point $([(n - p - q)/2] + 1)/n \approx 50\%$.

Remark: In the case $q = 1$ the proof of Theorem 1 becomes much easier and yields the following result for the breakdown point of the LTS estimator.

Corollary 1. *Denote $k'(Z_n)$ the maximal number of $x_j \in \{x_i; i = 1, \ldots, n\}$ lying on a hyperplane of $I\!R^p$. Then for any dataset $Z_n \subset I\!R^{p+1}$ with $k'(Z_n) < h$ it holds that*

$$\varepsilon_n^*(\hat{\mathcal{B}}_{LTS}, Z_n) = \frac{\min(n - h + 1, h - k'(Z_n))}{n}. \tag{3.2}$$

If Z_n comes from a continuous distribution F then almost surely $k'(Z_n) = p - 1$ yielding $\varepsilon_n^*(\hat{\mathcal{B}}_{LTS}, Z_n) = \min(n - h + 1, h - p + 1)/n$, as was already obtained by Hössjer (1994). In this case any $[(n + p)/2] \leq h \leq [(n + p + 1)/2]$ gives the maximal breakdown point.

4 Algorithm

Recently, Rousseeuw and Van Driessen (1999) developed a fast algorithm to compute the MCD location and scatter estimator. The basic tool for this algorithm was the so called C-step which guaranteed to decrease the MCD objective function. Similarly, the following theorem gives a C-step which can only decrease the MLTS objective function.

Theorem 2. *Take $H_1 \in \mathcal{H}$ with corresponding least squares estimates $\hat{\mathcal{B}}_1 := \hat{\mathcal{B}}_{LS}(H_1)$ and $\hat{\Sigma}_1 := \hat{\Sigma}_{LS}(H_1)$. If $\det(\hat{\Sigma}_1) > 0$ then denote by H_2 the set of indices of the observations corresponding with the h smallest residual distances $d_{1:n}(\hat{\mathcal{B}}_1, \hat{\Sigma}_1) \leq \cdots \leq d_{h:n}(\hat{\mathcal{B}}_1, \hat{\Sigma}_1)$. For $\hat{\mathcal{B}}_2 := \hat{\mathcal{B}}_{LS}(H_2)$ and $\hat{\Sigma}_2 := \hat{\Sigma}_{LS}(H_2)$, we have*

$$\det(\hat{\Sigma}_2) \leq \det(\hat{\Sigma}_1)$$

with equality if and only if $\hat{\mathcal{B}}_2 = \hat{\mathcal{B}}_1$ and $\hat{\Sigma}_2 = \hat{\Sigma}_1$.

Constructing in this way from H_1 a new subsample H_2 is called a C-step where, following Rousseeuw and Van Driessen (1999), C stands for "concentration" because the new subsample H_2 is more concentrated than H_1 in the sense that $\det(\hat{\Sigma}_2)$ is lower than $\det(\hat{\Sigma}_1)$.

The C-step of Theorem 2 forms the basis of our MLTS algorithm we will describe now. We start by drawing m random $p + q$ subsets J_m of $\{1, \ldots, n\}$ and compute the corresponding least squares estimators $\hat{\mathcal{B}}_m := \hat{\mathcal{B}}_{LS}(J_m)$ and $\hat{\Sigma}_m := \hat{\Sigma}_{LS}(J_m)$. If $\det(\hat{\Sigma}_m) = 0$ for some subset J_m then we draw additional points until $\det(\hat{\Sigma}_m) > 0$ or $\#J_m = h$. For each subset we compute the residual distances $d_i(\hat{\mathcal{B}}_m, \hat{\Sigma}_m)$ for $i = 1, \ldots, n$ and denote H_1 the subset corresponding to the h observations with smallest residual distances. Then we apply some

C-steps (e.g. two), lowering each time the value of the objective function. We then select the 10 subsets J_m which yielded the lowest determinants and for them we carry out further C-steps until convergence. The resulting subsample with lowest determinant among the 10 will be the final solution reported by the algorithm. For large datasets the algorithm can be speed up by using nested extensions as proposed by Rousseeuw and Van Driessen (1999).

5 The Functional

The functional form of the MLTS estimator can be defined as follows. Let K be an arbitrary $(p+q)$ dimensional distribution which represents the joint distribution of the carriers and response variables. Denote by $0 < \alpha < 1$ the mass not determining the MLTS estimator and define

$$\mathcal{D}_K(\alpha) = \{A|\, A \subset I\!R^{p+q} \text{ measurable and bounded with } P_K(A) = 1-\alpha\}. \tag{5.1}$$

To define the MLTS estimator at the distribution K we require that

$$P_K(\beta^t x = 0) < 1-\alpha \text{ for all } \beta \in I\!R^p \setminus \{0\}. \tag{5.2}$$

For each $A \in \mathcal{D}_K(\alpha)$, the least squares solution over the set A is then given by

$$\mathcal{B}_A(K) = \left(\int_A xx^t\, dK(x,y)\right)^{-1} \int_A xy^t\, dK(x,y) \tag{5.3}$$

and

$$\Sigma_A(K) = \frac{\int_A (y - \mathcal{B}_A(K)^t x)(y - \mathcal{B}_A(K)^t x)^t\, dK(x,y)}{1-\alpha}. \tag{5.4}$$

Furthermore, a set $\hat{A} \in \mathcal{D}_K(\alpha)$ is called an MLTS solution if $\det(\Sigma_{\hat{A}}(K)) \leq \det(\Sigma_A(K))$ for any other $A \in \mathcal{D}_K(\alpha)$. The MLTS functionals at the distribution K are then defined as

$$\mathcal{B}_{MLTS}(K) = \mathcal{B}_{\hat{A}}(K) \text{ and } \Sigma_{MLTS}(K) = c_\alpha \Sigma_{\hat{A}}(K). \tag{5.5}$$

The constant c_α can be chosen such that consistency will be obtained at the specified model. If the distribution K is not continuous, then the definition of $\mathcal{D}_K(\alpha)$ can be modified as in Croux and Haesbroeck (1999) to ensure that the set $\mathcal{D}_K(\alpha)$ is non-empty.

Now consider the multivariate regression model

$$y = \mathcal{B}^t x + \varepsilon$$

where $x = (x_1, \dots, x_p)$ is the p-dimensional vector of explanatory variables, $y = (y_1, \dots, y_q)$ is the q-dimensional vector of response variables and ε is the error term. We suppose that ε is independent of x and has a distribution F_Σ with density

$$f_\Sigma(u) = \frac{g(u^t \Sigma^{-1} u)}{\sqrt{\det(\Sigma)}}$$

where $\Sigma \in PDS(q)$. The function g is assumed to have a strictly negative derivative g' such that F_Σ is a unimodal elliptically symmetric distribution around the origin. The distribution of $z = (x, y)$ is denoted by H. A regularity condition (to avoid degenerate situations) on the model distribution H is that

$$P_H(\beta^t x + \gamma^t y = 0) < 1 - \alpha \tag{5.6}$$

for all $\beta \in I\!R^p$ and $\gamma \in I\!R^q$ not both equal to zero at the same time. If $\alpha = 0$ this regularity condition means that the distribution H is not completely concentrated on a $(p + q - 1)$-dimensional hyperplane. If $\alpha > 0$ this general position condition says that the maximal amount of probability mass of H lying on the same hyperplane must be lower than $1 - \alpha$. We first give the following proposition which says that the MLTS solution can always be taken as a cylinder.

Lemma 2. *Consider a distribution H satisfying* (5.6) *and an MLTS solution $\hat{A} \in D_H(\alpha)$. For any $(x, y) \in I\!R^{p+q}$ denote $d^2(x, y) = (y - \mathcal{B}_{\hat{A}}(H)^t x)^t (\Sigma_{\hat{A}}(H))^{-1} (y - \mathcal{B}_{\hat{A}}(H)^t x)$. Define the cylinder $\mathcal{E} = \{(x, y) \in I\!R^{p+q}; d^2(x, y) \le D_\alpha^2\}$ where D_α^2 is chosen such that $P_H(\mathcal{E}) = 1 - \alpha$. Then it holds that*

$$\mathcal{B}_{\mathcal{E}}(H) = \mathcal{B}_{\hat{A}}(H) \text{ and } \Sigma_{\mathcal{E}}(H) = \Sigma_{\hat{A}}(H).$$

We now show that the functionals $\mathcal{B}_{MLTS}(H)$ and $\Sigma_{MLTS}(H)$ defined by (5.5) for some well chosen constant c_α are Fisher-consistent for the parameters $\mathcal{B}$ and Σ.

Theorem 3. *Denote*

$$c_\alpha = \frac{1 - \alpha}{\int_{\|u\|^2 \le q_\alpha} u_1^2 \, dF_0(u)}$$

where $F_0 = F_{I_q}$ is the central error distribution and $q_\alpha = K^{-1}(1 - \alpha)$ with $K(t) = P_{F_0}(U^t U \le t)$. Then the functionals $\mathcal{B}_{MLTS}$ and Σ_{MLTS} are Fisher-consistent estimators for the parameters $\mathcal{B}$ and Σ at the model distribution H:

$$\mathcal{B}_{MLTS}(H) = \mathcal{B} \quad \text{and} \quad \Sigma_{MLTS}(H) = \Sigma.$$

Note that for obtaining the above consistency result we only made an assumption on the distribution of the errors, but not on the distribution of (x, y). For multivariate normal errors we can take $c_\alpha = (1-\alpha)/F_{\chi^2_{p+2}}(q_\alpha)$ with $q_\alpha = \chi^2_{p,1-\alpha}$, the upper α percent point of the χ^2_p distribution.

6 The influence function and asymptotic variances

The influence function of a functional T at the distribution H measures the effect on T of adding a small mass at $z = (x, y)$. If we denote the point mass at z by Δ_z and consider the contaminated distribution $H_{\varepsilon,z} = (1-\varepsilon)H + \varepsilon\Delta_z$ then the influence function is given by

$$IF(z; T, H) = \lim_{\varepsilon \downarrow 0} \frac{T(H_{\varepsilon,z}) - T(H)}{\varepsilon} = \frac{\partial}{\partial\varepsilon} T(H_{\varepsilon,z})\Big|_{\varepsilon=0}.$$

(See Hampel et al. 1986.) It can easily be seen that the MLTS is equivariant for affine transformations of the regressors and responses and for regression transformations which add a linear function of the explanatory variables to the responses. Therefore, it suffices to derive the influence function at a model distribution H_0 for which $\mathcal{B} = 0$ and the error distribution $F_0 = F_{I_q}$ with density $f_0(y) = g(y^t y)$. The following theorem gives the influence function of the MLTS regression functional at H_0.

Theorem 4. *With the notations from above, we have that*

$$IF(z; \mathcal{B}_{MLTS}, H_0) = E_{H_0}[xx^t]^{-1} \frac{xy^t}{-2c_2} I(\|y\|^2 \leq q_\alpha) \tag{6.1}$$

where c_2 is given by

$$c_2 = \frac{\pi^{q/2}}{\Gamma(q/2+1)} \int_0^{\sqrt{q_\alpha}} r^{q+1} g'(r^2)\, dr$$

Note that the influence function is bounded in y but unbounded in x. Closer inspection of (6.1) shows, however, that only good leverage points, which have outlying x but satisfy the regression model, can have a high effect on the MLTS estimator. Bad leverage points will give a zero influence. In the case of simple regression, the influence function of the LTS slope has been plotted in Croux et al. (1994, Figure 3d).

Remark 1: The influence function of the MCD location estimator T_q at a q-dimensional spherical distribution F_0 can be obtained from Butler, Davies and Jhun (1993) or Croux and Haesbroeck (1999). With the notations as before it is given by

$$IF(y, T_q, F_0) = \frac{y}{-2c_2} I(\|y\|^2 \leq q_\alpha).$$

Therefore, it follows that the influence function of $\mathcal{B}_{MLTS}$ can be rewritten as

$$IF(z;\mathcal{B}_{MLTS},H_0)=E_{H_0}[xx^t]^{-1}\,x\,IF(y;T_q,F_0)^t. \tag{6.2}$$

Remark 2: In the case $q=1$ we have $c_2=\int_{-\sqrt{q_\alpha}}^{\sqrt{q_\alpha}}g'(y^2)y^2\,dy=\sqrt{q_\alpha}f(\sqrt{q_\alpha})-((1-\alpha)/2)$ so we obtain

$$IF(z;\mathcal{B}_{MLTS},H_0)=E_{H_0}[xx^t]^{-1}\frac{xyI(y^2\le q_\alpha)}{1-\alpha-2\sqrt{q_\alpha}f(\sqrt{q_\alpha})}$$

which is the expression for the influence function of the LTS estimator.

Remark 3: Similarly as in Theorem 4 it can be shown that

$$IF(z;\Sigma_{MLTS},H_0)=IF(y,C_q,F_0)$$

where C_q is the q-dimensional MCD scatter estimator. The influence function of the MCD scatter estimator at elliptical distributions can be obtained from Croux and Haesbroeck (1999).

The asymptotic variance-covariance matrix of $\mathcal{B}_{MLTS}$ can now be computed by means of $ASV(\mathcal{B}_{MLTS},H_0)=E_H[IF(z;\mathcal{B}_{MLTS},H_0)\otimes IF(z;\mathcal{B}_{MLTS},H_0)^t]$ (see e.g. Hampel et al. 1986). Here $A\otimes B$ denotes the Kronecker product of a $(d_1\times d_2)$ matrix A with a $(d_3\times d_4)$ matrix B, which results in a $(d_1d_3\times d_2d_4)$ matrix with d_1d_2 blocks of size $(d_3\times d_4)$. For $1\le j\le d_1$ and $1\le k\le d_2$ the (j,k)-th block equals $a_{jk}B$, where a_{jk} are the elements of the matrix A. Let us denote $\Sigma_x:=E_{H_0}[xx^t]$, then expression (6.2) implies that

$$ASV(\mathcal{B}_{MLTS},H_0)=D_{p,q}(\mathrm{diag}(ASV(T_q,F_0))\otimes\Sigma_x^{-1}) \tag{6.3}$$

where the commutation matrix $D_{p,q}$ is a $(pq\times pq)$ matrix consisting of pq blocks of size $(q\times p)$. For $1\le l\le p$ and $1\le m\le q$ the (l,m)th block of $D_{p,q}$ equals the $(q\times p)$ matrix Δ_{ml} which is 1 at entry (m,l) and 0 everywhere else.

From (6.3) it follows that for every $1\le i\le p$ and $1\le j\le q$ the asymptotic covariance matrix of $(\mathcal{B}_{MLTS})_{ij}$ is given by $\Delta_{ji}\Sigma_x^{-1}ASV((T_q)_j,F_0))$ which implies that the asymptotic variance of $(\mathcal{B}_{MLTS})_{ij}$ equals

$$ASV((\mathcal{B}_{MLTS})_{ij},H_0)=E_H[IF^2(z;(\mathcal{B}_{MLTS})_{ij},H_0)]=(\Sigma_x^{-1})_{ii}ASV((T_q)_j,F_0).$$

For $i\neq i'$ we obtain the asymptotic covariances

$$\begin{aligned}ASC((\mathcal{B}_{MLTS})_{ij},(\mathcal{B}_{MLTS})_{i'j},H_0)&=E_H[IF(z;(\mathcal{B}_{MLTS})_{ij},H_0)IF(z;(\mathcal{B}_{MLTS})_{i'j},H_0)]\\&=(\Sigma_x^{-1})_{ii'}ASV((T_q)_j,F_0)\end{aligned}$$

Table 1: Asymptotic relative efficiency of the MLTS estimator w.r.t. the classical estimator at the normal distribution for several values of q.

α	$q=2$	$q=3$	$q=5$	$q=10$	$q=30$
0.25	0.403	0.466	0.531	0.597	0.664
0.5	0.153	0.204	0.262	0.327	0.398

and all other asymptotic covariances (for $j' \neq j$) equal 0.

Due to affine equivariance, we may consider w.l.o.g. the case where $\Sigma_x = I_p$. Then all asymptotic covariances are zero, while $ASV((\mathcal{B}_{MLTS})_{ij}, H_0) = ASV((T_q)_j, F_0)$ for all $1 \leq i \leq p$ and $1 \leq j \leq q$. The limit case $\alpha = 0$ yields the asymptotic variance of the least squares estimator $ASV((\mathcal{B}_{LS})_{ij}, H_0) = ASV(M_j, F_0)$ where M is the functional form of the sample mean. Therefore, we can compute the asymptotic relative efficiency of the MLTS estimator at the model distribution H_0 with respect to the least squares estimator as

$$ARE((\mathcal{B}_{MLTS})_{ij}, H_0) = \frac{ASV((\mathcal{B}_{LS})_{ij}, H_0)}{ASV((\mathcal{B}_{MLTS})_{ij}, H_0)} = \frac{ASV(M_j, F_0)}{ASV((T_q)_j, F_0)} = ARE((T_q)_j, F_0)$$

for all $1 \leq i \leq p$ and $1 \leq j \leq q$. Hence the asymptotic relative efficiency of the MLTS estimator in $p+q$ dimensions does not depend on the distribution of the carriers, but only on the distribution of the errors and equals the asymptotic relative efficiency of the q-dimensional MCD location estimator at the error distribution F_0. For the normal distribution these relative efficiencies are given in Table 1. Note that the efficiency of MLTS does not depend on p, the number of explanatory variables, but only on the number of dependent variables.

7 Reweighting and one-step improvements

The efficiency of MLTS can be quite low as can be seen from Table 1. Therefore, we now introduce some methods that improve the performance of the MLTS.

One way to increase the efficiency of the MLTS is to consider the one-step reweighted MLTS. If $\hat{\mathcal{B}}_{MLTS}$ and $\hat{\Sigma}_{MLTS}$ denote the initial MLTS estimates, then the one-step reweighted MLTS estimates (RMLTS) are defined as

$$\hat{\mathcal{B}}_{RMLTS} := \hat{\mathcal{B}}_{LS}(J) \quad \text{and} \quad \hat{\Sigma}_{RMLTS} := c_\delta \operatorname{cov}(J, \hat{\mathcal{B}}_{LS}(J)),$$

where $J = \{j : d_j^2(\hat{\mathcal{B}}_{MLTS}, \hat{\Sigma}_{MLTS}) \leq q_\delta\}$. Here δ is the trimming fraction and $c_\delta := (1-\delta)/\int_{\|u\|^2 \leq q_\delta} u_1^2 \, dF_0(u)$ a consistency factor to obtain Fisher-consistency at the model distribution. Following Rousseeuw and Leroy (1987) we used $\delta = 0.01$ and $q_\delta = \chi^2_{q,1-\delta}$ the corresponding quantile of the χ^2 distribution with q degrees of freedom. In the case of multivariate normal errors we have $c_\delta = (1-\delta)/F_{\chi^2_{p+2}}(q_\delta)$.

It has been shown that one-step GM estimators are highly efficient robust estimators for univariate linear regression (see e.g. Simpson et al. 1992, Coakley and Hettmansperger 1993). Therefore, as an alternative for the RMLTS, we also construct a multivariate generalization of one-step GM estimators that use MLTS as initial estimator. With $\hat{\mathcal{B}}_{MLTS}$ and $\hat{\Sigma}_{MLTS}$ the initial MLTS estimates, the multivariate generalization of the one-step GM estimators is given by

$$\hat{\mathcal{B}}^1 = \hat{\mathcal{B}}_{MLTS} + (X^tVX)^{-1}X^tW\tilde{R}.$$

The diagonal matrix $W = \text{diag}(w_i)$ only depends on the explanatory variables x_i. Following Simpson et al. (1992), for a model with intercept we put $w_i = w(x_i) = \min(1, \frac{\chi^2_{p-1,0.95}}{h(x_i)^2})$ where $h(x_i)$ is the robust distance of x_i based on the MCD mean $T_{p-1}(X)$ and scatter $C_{p-1}(X)$ of the explanatory variables, given by

$$h(x_i) := \sqrt{(x_i - T_{p-1}(X))^t(C_{p-1}(X))^{-1}(x_i - T_{p-1}(X))}.$$

The diagonal matrix $V = \text{diag}(v_i)$ depends on the robust distances of the residuals $d_i(\hat{\mathcal{B}}_{MLTS}, \hat{\Sigma}_{MLTS})$ and the weights w_i. The diagonal elements are given by $v_i = w_i^{1-a}\psi'(d_i(\hat{\mathcal{B}}_{MLTS}, \hat{\Sigma}_{MLTS})/w_i^a)$. Finally, the matrix $\tilde{R} = (\tilde{r}_1, \ldots, \tilde{r}_n)^t$ is an adjusted residual matrix whose elements are given by $\tilde{r}_i := \psi(d_i(\hat{\mathcal{B}}_{MLTS}, \hat{\Sigma}_{MLTS})/w_i^a)r_i(\hat{\mathcal{B}}_{MLTS})/d_i(\hat{\mathcal{B}}_{MLTS}, \hat{\Sigma}_{MLTS})$.

We will consider the choices $a = 0$ and $a = 1$ which correspond to the Mallows and Schweppe type one-step M-estimators respectively. Simpson et al. (1992) showed that using Mallows weights and Hampel's three part redescending psi function yields a robust, locally stable one-step M-estimator. In the multivariate setting we use Hampel's psi function with constants $(a, b, c) = (\sqrt{\chi^2_{q,0.80}}, \sqrt{\chi^2_{q,0.997}}, 10)$.

To obtain a highly efficient estimator Coakley and Hettmansperger (1993) proposed to use Schweppe type weights and the Huber psi function $\psi_k(t) = \min(|t|, k)\,\text{sign}(t)$. The constant k is the cutoff point for outliers which we set equal to $k = \sqrt{\chi^2_{q,0.80}}$. From now on, the multivariate generalizations of the Mallows and Schweppe one-step M-estimators will be denoted as MM1M and MS1M respectively.

8 Finite-sample simulations

8.1 Finite-sample performance

In this section we investigate the finite-sample performance of the MLTS estimator. Therefore, we will compare the asymptotic efficiency obtained in the previous section with finite-sample efficiencies obtained by simulation. To this end, we performed the following simulations. For various sample sizes n, and for $p = 3$ and $q = 3$, we generated $m = 1000$ regression datasets of size n. The response variables were generated from the multivariate standard normal distribution $N(0, I_q)$, and w.l.o.g. we took $\mathcal{B} = 0$ in the multivariate regression model. We set the pth regressor equal to one, so we consider a regression model with intercept. The remaining $p - 1$ explanatory variables were generated from the following distributions:

1. (NOR) The multivariate standard normal distribution $N(0, I_{p-1})$.

2. (EXP) The distribution of $U = V - 1$, where V is a vector of $p-1$ independent variables and each variable follows an exponential distribution with mean one.

3. (CAU) The multivariate Cauchy which is defined as the distribution of $(\sqrt{V})^{-1}U$, where $U \sim N(0, I_{p-1})$ is independent of $V \sim \chi_1^2$. (See e.g. Johnson and Kotz 1972, p. 134.)

In this simulation setup, the last row of $\mathcal{B}$ is the intercept vector and the matrix formed by the $p - 1$ first rows of $\mathcal{B}$, which we will denote by $\mathcal{B}^0$, is the slope matrix. For the subset size h, we considered two typical choices, namely, $h = [(n + p + q + 1)/2]$ (corresponding to $\alpha = 0.5$) which yields the highest breakdown point and $h \approx 0.75n$ (corresponding to $\alpha = 0.25$) which gives a better compromise between breakdown and efficiency.

For each simulated dataset $Z^{(l)}, l = 1, \ldots, m$ we computed the $(p \times q)$ regression matrix $\hat{\mathcal{B}}^{(l)}_{MLTS}$. The Monte Carlo variance of a regression coefficient $(\hat{\mathcal{B}}_{MLTS})_{jk}$ is measured as $\text{Var}((\hat{\mathcal{B}}_{MLTS})_{jk}) = n \underset{l}{\text{var}}((\hat{\mathcal{B}}_{MLTS})^{(l)}_{jk})$ for $j = 1, \ldots, p$ and $k = 1, \ldots, q$. The variance of the estimated slope matrix $\hat{\mathcal{B}}^0_{MLTS}$ is then summarized by $\underset{j,k}{\text{ave}}(\text{Var}((\hat{\mathcal{B}}_{MLTS})_{jk}))$ for $1 \leq j \leq p-1$ and $1 \leq k \leq q$ while its inverse measures the finite-sample efficiency of the slope. Similarly we computed the finite-sample efficiency of the intercept vector.

Table 2 shows the finite-sample efficiencies of the MLTS estimator obtained by simulation for sample size n equal to 100, 300, and 500. We see that the finite-sample efficiencies of the MLTS converge to the corresponding asymptotic efficiencies which are listed under $n = \infty$

Table 2: Finite-sample efficiencies of the MLTS slope and intercept at normal (NOR) or exponential (EXP) carrier distributions and normal error distribution for $p = 3, q = 3$ and several values of the sample size n.

		$n = 100$		$n = 300$		$n = 500$		$n = \infty$
		NOR	EXP	NOR	EXP	NOR	EXP	
$\alpha = 0.50$	slope	0.250	0.208	0.221	0.208	0.213	0.206	0.204
	intercept	0.249	0.241	0.232	0.232	0.220	0.221	0.204
$\alpha = 0.25$	slope	0.480	0.437	0.477	0.462	0.470	0.458	0.466
	intercept	0.464	0.464	0.482	0.484	0.469	0.471	0.466

in Table 2. Note that efficiencies for $\alpha = 0.25$ are always higher than the corresponding efficiency for $\alpha = 0.5$. From Table 2 we also see that results obtained for the asymmetric exponential carriers are comparable to those obtained for normal carriers. This confirms that the efficiency of MLTS does not depend on the distribution of the carriers when the carriers are uncorrelated. Results for Cauchy carriers are omitted because in this case the asymptotic variance of the MLTS and LS estimators do not exist.

To compare the performance of the multivariate one-step M-estimators and RMLTS with that of MLTS we used the same simulation setup as before and for each of the estimators we computed the mean squared error of the slope matrix and intercept vector. For a univariate estimator T, the mean squared error is given by

$$\mathrm{MSE}(T) = n \operatorname*{ave}_{l} (T^{(l)} - \theta)^2$$

where θ is the true value of the parameter. The MSE of the MLTS slope matrix $\hat{\mathcal{B}}^0_{MLTS}$ is then defined as

$$\mathrm{MSE}(\hat{\mathcal{B}}^0_{MLTS}) = \underset{1 \leq j \leq p-1, 1 \leq k \leq q}{ave} (\mathrm{MSE}((\hat{\mathcal{B}}_{MLTS})_{jk}))$$

and similarly for the intercept vector. The MSE of the slope matrix and intercept vector of the RMLTS and one-step M-estimators are computed analogously.

For data generated from the normal (NOR) distribution, Figure 1a shows the resulting MSE for the slope of the 25% breakdown MLTS (MLTS25), reweighted MLTS (RMLTS25) and Mallows (MM1M25) and Schweppe (MS1M25) type one-step M-estimators. The MSE of the 50% breakdown estimators is shown in Figure 1b. Throughout this section the results for the slope will be shown and the results for the intercept will be omitted because they yielded the same conclusions. From Figure 1 we see that all three proposals clearly improve

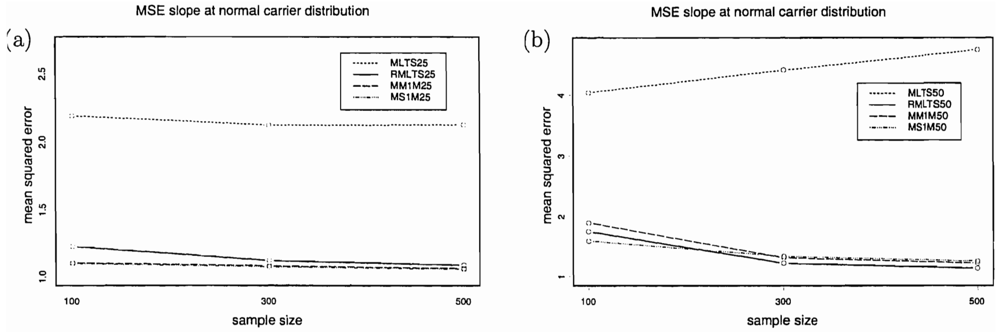


Figure 1: MSE at the normal distribution for the MLTS, the one-step reweighted MLTS (RMLTS) and Mallows (MM1M) and Schweppe (MS1M) type one-step GM estimators. (a) 25% breakdown point; (b) 50% breakdown point.

the performance of the initial MLTS estimator. Moreover, the MSE of the one-step M-estimators is comparable or slightly better than the MSE of the corresponding RMLTS estimator.

In Figure 2 we investigate the performance of the estimators at asymmetric (EXP) and long tailed carrier (CAU) distributions. The MSE of the MLTS, RMLTS, MM1M and MS1M estimators for carriers generated from the exponential distribution (EXP) are shown in Figure 2a and Figure 2b. Figure 2c and Figure 2d show the results for carriers from the Cauchy distribution (CAU). From these plots we see that the RMLTS in all cases improves the performance of the initial MLTS estimator. On the other hand, the MM1M estimator improves the MSE of the initial MLTS at exponential carrier distributions but yields a much worse MSE for Cauchy carrier distributions. Finally, in all cases the MSE of the MS1M estimator is comparable or much worse than the MSE of the initial MLTS. Hence, the one-step M-estimators only work well for normal carrier distributions. In general, we conclude that overall the RMLTS has the best performance.

8.2 Finite-sample robustness

To study the finite-sample robustness of the MLTS estimator we carried out simulations with contaminated datasets. We consider the following types of outliers: an observation $z_i = (x_i, y_i)$ which does not follow the linear pattern of the majority of the data, but whose x_i is not outlying, is called a vertical outlier. A data point whose x_i is outlying is called a

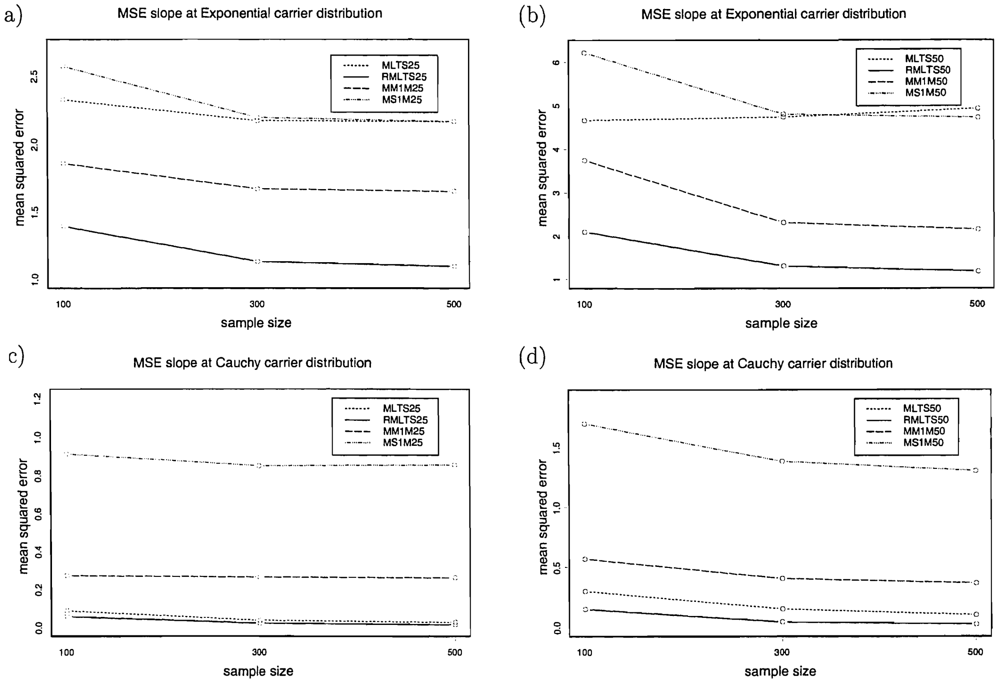


Figure 2: MSE at the Exponential and Cauchy carrier distribution for the MLTS, RMLTS, MM1M and MS1M estimators with 25% and 50% breakdown point.

leverage point. We say that such a data point is a bad leverage point when it does not follow the linear pattern of the majority, otherwise it is called a good leverage point (which does not harm the fit). Since regression estimators often break down in the presence of vertical outliers or bad leverage points, we generated datasets with these two types of outliers.

To generate contaminated datasets with vertical outliers we started from the uncontaminated datasets as before and then we replaced 20% of the responses y_i by q response variables distributed according to $N(5\sqrt{\chi^2_{q,.99}}, 1.5)$. Figure 3 shows the MSE for respectively normal, Exponential, and Cauchy carrier distributions. From these plots we see that MLTS and RMLTS always have a low MSE in the presence of vertical outliers which confirms that these estimators are robust to vertical outliers. Furthermore, in all cases RMLTS improves the MSE of the initial MLTS and in most cases this improvement is substantial. On the other hand we see that the MM1M is comparable to RMLTS in the case of normal carriers.

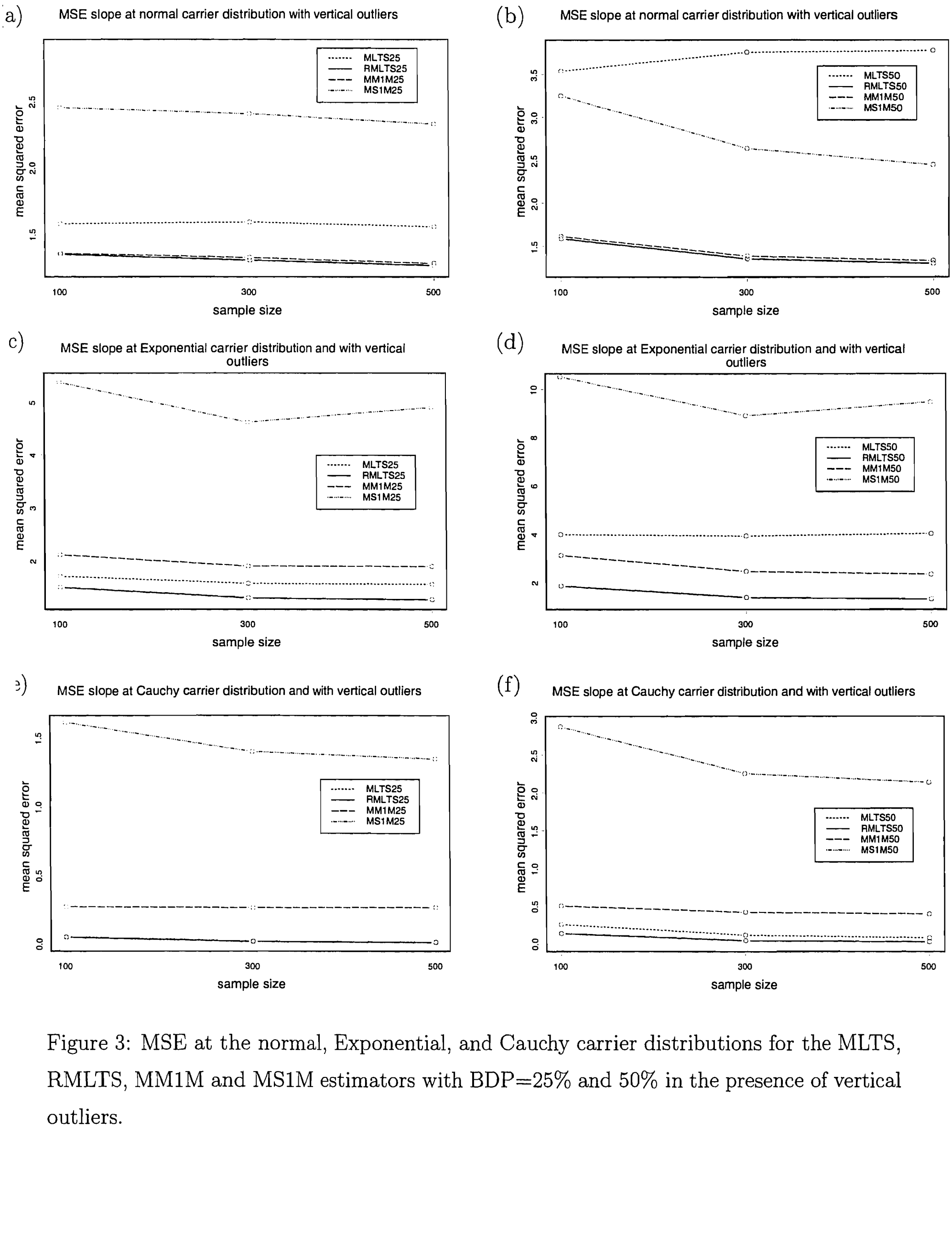
It still improves the MSE of the initial MLTS in the case of exponential carriers and 50% breakdown point, but in the other situations it is worse than the initial MLTS. Finally, the MS1M in almost all cases gives a much worse result than the initial MLTS which shows that the gain in efficiency obtained by MS1M leads to an increased bias in the presence of outliers.

To generate contaminated datasets with bad leverage points we started from the uncontaminated datasets as before and then we replaced 20% of the data with observations for which the $p-1$ independent variables were generated according to $N(5\sqrt{\chi^2_{p-1,.99}}, 1.5)$ and the q dependent variables were generated from $N(5\sqrt{\chi^2_{q,.99}}, 1.5)$. Figure 4 shows the MSE for respectively normal, Exponential, and Cauchy carrier distributions. From these plots we see that MLTS and RMLTS always have a low MSE in the presence of bad leverage points, hence these estimators are also robust to bad leverage points. As before, we also have that in all cases RMLTS improves the MSE of the initial MLTS. On the other hand, the MM1M improves the MSE of the initial MLTS in case of normal or exponential carriers, and can even be better than RMLTS, but it is much worse for the Cauchy carrier distribution. Finally, as with vertical outliers, in most cases the MSE of MS1M is much worse than the MSE of the initial MLTS. Note that for Cauchy carrier distributions we have omitted the MSE of MS1M because it was even much bigger than the MSE of the initial MLTS.

To summarize, our simulations with contaminated datasets confirmed that vertical outliers and bad leverage points have a small influence on the MLTS and RMLTS estimators, thus MLTS and RMLTS are robust to vertical outliers as well as bad leverage points. In all cases the RMLTS improved the result of the initial MLTS. On the other hand we noted that outliers can have a much higher influence on the one-step M-estimators which can perform much worse than the initial MLTS in the presence of outliers.

9 Example

To illustrate the MLTS method in practice, we use a real dataset of Charnes et al. (1981). This dataset consists of 70 observations on 5 explicative variables and 3 response variables. For students of 70 school sites in the U.S. the following five inputs were measured: education level of mother (x_1), highest occupation of a family member (x_2), number of parental visits to the school (x_3), parent counseling concerning school-related topics (x_4), and the number of teachers at the school site (x_5). The three outputs are the total reading score measured


MSE slope at normal carrier distribution with vertical outliers

MSE slope at normal carrier distribution with vertical outliers

MSE slope at Exponential carrier distribution and with vertical outliers

MSE slope at Exponential carrier distribution and with vertical outliers

MSE slope at Cauchy carrier distribution and with vertical outliers

MSE slope at Cauchy carrier distribution and with vertical outliers

Figure 3: MSE at the normal, Exponential, and Cauchy carrier distributions for the MLTS, RMLTS, MM1M and MS1M estimators with BDP=25% and 50% in the presence of vertical outliers.

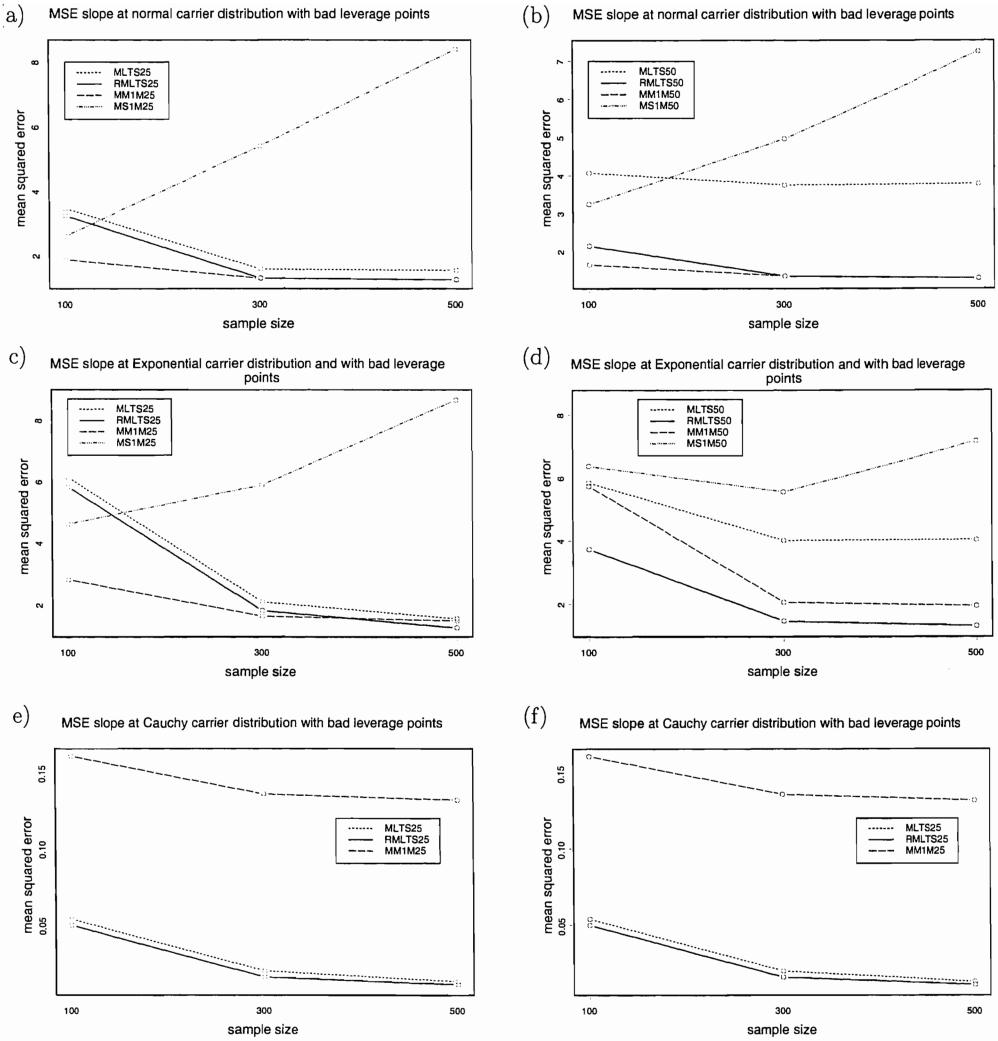


Figure 4: MSE at the normal, Exponential, and Cauchy carrier distributions for the MLTS, RMLTS, MM1M, and MS1M estimators with BDP=25% and 50% in the presence of bad leverage points.

Table 3: Estimates of the regression parameters for the school data obtained by RMLTS with 25% BPD and trimming proportion $\delta = 0.01$ and by least squares. the last column are the intercept estimates.

RMLTS-estimator					
0.098	4.760	0.087	-0.750	-0.171	1.998
0.031	5.146	0.115	-0.713	-0.228	2.616
-0.013	1.575	0.270	0.003	0.035	0.176
LS-estimator					
0.203	3.745	-0.283	-0.091	-0.181	-0.179
0.110	4.770	-0.529	0.143	-0.341	-0.366
-0.045	2.227	0.195	-0.056	0.011	-0.041

by the Metropolitan Achievement Test (y_1), the total mathematics score measured by the Metropolitan Achievement Test (y_2), and the Coopersmith self-esteem inventory (y_3). We consider a multivariate regression model with intercept, hence $p = 6$ and $q = 3$. We applied the one-step reweighted MLTS regression with $\alpha = 0.25$ and $\delta = 0.99$ to these school data and we denote the resulting fit as $\hat{\mathcal{B}}_{RMLTS}$. The estimate for the covariance of the errors is denoted as $\hat{\Sigma}_{RMLTS}$. The estimates of the regression coefficients are reported in Table 3 together with the classical least squares estimates. We see that there are differences both in magnitude and sign between the RMLTS and LS estimates indicating the presence of outliers that influenced the classical estimates.

In order to detect outliers in multivariate linear regression we construct the following diagnostic plot. First we compute the robust distances $d_i(\hat{\mathcal{B}}_{RMLTS}, \hat{\Sigma}_{RMLTS})$ in the q-dimensional residual space. Then we compute the one-step reweighted MCD mean $T^1_{p-1}(X)$ and scatter $C^1_{p-1}(X)$ of the explanatory variables and obtain the corresponding robust distances $h(x_i)$ in the space of explicative variables. Since, for outlier-free samples, $h(x_i)^2$ and $d_i^2(\hat{\mathcal{B}}_{RMLTS}, \hat{\Sigma}_{RMLTS})$ roughly have chi-squared distributions with $p-1$ and q d.f., respectively, we can classify the ith observation as a *high leverage point* if $h(x_i)^2 > \chi^2_{\delta,p-1}$, and as a *multivariate regression outlier* if $d_i^2(\hat{\mathcal{B}}_{RMLTS}, \hat{\Sigma}_{RMLTS}) > \chi^2_{\delta,q}$ where $\chi^2_{\delta,r}$ denotes the δ quantile of a Chi-square distribution with r d.f. Plotting the robust residual distances $d_i(\hat{\mathcal{B}}_{RMLTS}, \hat{\Sigma}_{RMLTS})$ versus the robust distances $h(x_i)$ of the x_i and drawing the cutoff lines $h = \sqrt{\chi^2_{\delta,p-1}}$ and $d = \sqrt{\chi^2_{\delta,q}}$ allows us to detect vertical outliers, good and bad leverage points. This plot is a generalization to multivariate regression of the diagnostic plot pro-

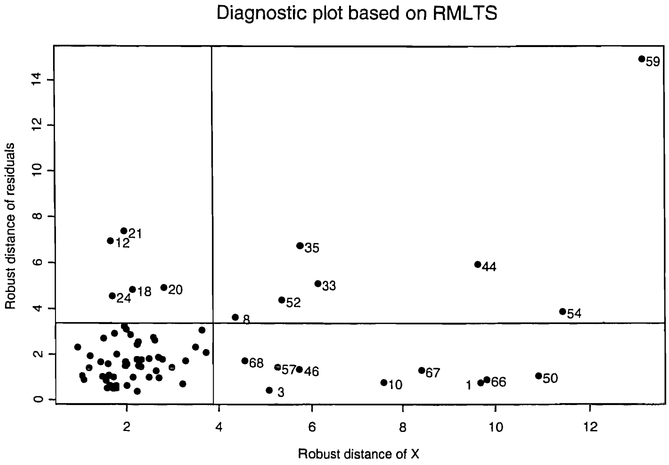


Figure 5: Diagnostic plot based on the one-step reweighted MLTS procedure for the school data.

posed by Rousseeuw and van Zomeren (1990) for univariate regression.

Figure 5 shows the diagnostic plot for the school data of Charnes et al. (1981) obtained with the one-step reweighted MLTS. This plot clearly shows that the dataset contains one very large bad leverage point (59). There are also two moderate to large bad leverage points (35,44) and two moderate to large vertical outliers (12 and 21). Moreover, there are at least five good leverage points (10,67,1,66,50). On the other hand, the diagnostic plot of the school data obtained by multivariate least squares regression in Figure 6 does not reveal any bad leverage points. Only a small vertical outlier and some small to moderate good leverage points are detected. This clearly indicates that the multivariate least squares estimator is attracted by the bad leverage points which leads to masking of the outliers in the dataset.

10 Conclusions

In this paper we have introduced the multivariate least trimmed squared estimator. We have given three equivalent definitions of the MLTS estimator which allow us to completely investigate and explain the behavior of the estimator. The MLTS has a positive breakdown point which depends on the subset size h to be chosen by the user. The choice of h is a trade-off between efficiency and breakdown. Two practical choices are $h = [(n+p+q+1)/2]$ which yields the maximal breakdown point $\varepsilon_n^* \approx 50\%$ and $h \approx 0.75n$ which gives a

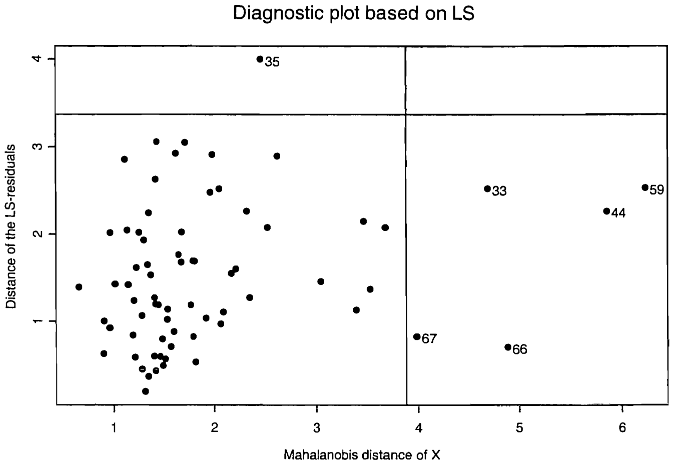


Figure 6: Diagnostic plot based on the LS procedure for the school data.

better compromise between breakdown (25%) and efficiency. We have defined the MLTS functional and shown that it is Fisher-consistent at the multivariate regression model with elliptically symmetric error distribution. Note that we did not make any hypothesis of symmetry on the distribution of the explanatory variables, we only assumed a regularity condition to avoid degenerate situations. The influence function and asymptotic variances of the MLTS functional have been derived. Since MLTS generalizes both LTS and MCD, these general results for MLTS close some gaps in the existing literature on LTS and MCD. For instance, a formal proof of the MCD breakdown point is now available. Based on a C-step theorem we have constructed a time-efficient algorithm to compute the MLTS estimator. This algorithm has been used to perform finite-sample simulations which investigate both efficiency and robustness. We also investigated the one-step reweighted MLTS estimator and multivariate one-step GM estimators based on MLTS. In all situations the reweighted MLTS improved the initial MLTS estimator. The one-step GM estimators improved the MLTS at the normal distribution, but simulations showed that these estimators can behave badly in other situations. Therefore, we recommend to use the one-step reweighted MLTS.

11 Appendix

Proof of Lemma 1. For ease of notation, let $\Sigma_{LS}(K) := \Sigma_{LS}$ and drop the subscript K. Put $u = \Sigma_{LS}^{-1/2}\varepsilon$. Then $E[(y-\mathcal{B}_{LS}^t x)^t \Sigma_{LS}^{-1}(y-\mathcal{B}_{LS}^t x)] = E[u^t u] = \operatorname{tr} E[uu^t] = \operatorname{tr}(\Sigma_{LS}^{-1/2} E[\varepsilon\varepsilon^t]\Sigma_{LS}^{-1/2}) =$

$\operatorname{tr} I_q = q$, so $(\mathcal{B}_{LS}, \Sigma_{LS})$ satisfies condition (2.4). Take any $b \in I\!R^{p\times q}$ and any Δ a positive definite symmetric matrix of size q such that (2.4) holds. There exists an orthogonal matrix P and $\lambda_1 \geq \cdots \geq \lambda_q > 0$ such that $\Delta = \Sigma_{LS}^{1/2} P \Lambda P^t \Sigma_{LS}^{1/2}$ where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_q)$. Put $v = P^t \Sigma_{LS}^{-1/2}(y - b^t x)$. Then we obtain

$$q = E[(y - b^t x)^t \Delta^{-1}(y - b^t x)] = E[v^t \Lambda^{-1} v] = \sum_{i=1}^{q} \lambda_i^{-1} E[v_i^2] \tag{11.1}$$

On the other hand, since $E[x\varepsilon^t] = 0$, we have that

$$\begin{aligned} E[vv^t] &= P^t \Sigma_{LS}^{-1/2} E[(\varepsilon + (\mathcal{B}_{LS} - b)^t x)(\varepsilon + (\mathcal{B}_{LS} - b)^t x)^t] \Sigma_{LS}^{-1/2} P \\ &= P^t (I_q + \Sigma_{LS}^{-1/2} (\mathcal{B}_{LS} - b)^t E[xx^t](\mathcal{B}_{LS} - b)\Sigma_{LS}^{-1/2}) P \\ &= I_q + ((\mathcal{B}_{LS} - b)\Sigma_{LS}^{-1/2} P)^t E[xx^t]((\mathcal{B}_{LS} - b)\Sigma_{LS}^{-1/2} P). \end{aligned} \tag{11.2}$$

Taking the diagonal elements of (11.2) and inserting them in (11.1) yields

$$q = \sum_{i=1}^{q} \lambda_i^{-1} + \sum_{i=1}^{q} \lambda_i^{-1} ((\mathcal{B}_{LS} - b)\Sigma_{LS}^{-1/2} P)_i^t E[xx^t]((\mathcal{B}_{LS} - b)\Sigma_{LS}^{-1/2} P)_i \geq \sum_{i=1}^{q} \lambda_i^{-1}, \tag{11.3}$$

with $((\mathcal{B}_{LS} - b)\Sigma_{LS}^{-1/2} P)_i$ the i-th column of this matrix. Furthermore, by definition of Δ and the relation between an arithmetic and geometric mean, we have

$$\sum_{i=1}^{q} \frac{1}{\lambda_i} \geq q(\prod_{i=1}^{q} \frac{1}{\lambda_i})^{1/q} = q(\det \Lambda)^{-1/q} = q(\frac{\det \Sigma_{LS}}{\det \Delta})^{1/q}. \tag{11.4}$$

From the last two inequalities (11.3) and (11.4) we see that $\det \Sigma_{LS} \leq \det \Delta$, showing already that $(\mathcal{B}_{LS}, \Sigma_{LS})$ solves the minimization problem.

Moreover, equality in (11.3) only occurs if all $((\mathcal{B}_{LS} - b)\Sigma_{LS}^{-1/2} P)_i = 0$, thus if $b = \mathcal{B}_{LS}$. In order to have $\det \Sigma_{LS} = \det \Delta$, also (11.4) needs to become an equality, which can only occur if all λ_i are equal to one, implying $\Delta = \Sigma_{LS}$. Hereby, we have also proved the uniqueness part. □

Proof of Proposition 1: Take $\hat{H} \in \underset{H}{\operatorname{argmin}} \det \hat{\Sigma}_{LS}(H)$. We first prove that $\hat{\mathcal{B}}_{LS}(\hat{H})$ minimizes $\det \mathrm{MCD}_q(\mathcal{B})$. Take $\mathcal{B} \in I\!R^{p\times q}$ arbitrarily, then by definition of the MCD there exists a $H \in \mathcal{H}$ such that $\mathrm{MCD}_q(\mathcal{B}) = \mathrm{Cov}_0(H, \mathcal{B})$. Using properties of traces, it follows that

$$\frac{1}{h} \sum_{j\in H} r_j(\mathcal{B})(\mathrm{Cov}_0(H, \mathcal{B}))^{-1} r_j(\mathcal{B})^t = q. \tag{11.5}$$

Since the data are in general position, Lemma 1 can be applied:

$$\det \mathrm{MCD}_q(\mathcal{B}) = \det \mathrm{Cov}_0(H, \mathcal{B}) \geq \det \hat{\Sigma}_{LS}(H) \geq \det \hat{\Sigma}_{LS}(\hat{H}) = \det \mathrm{Cov}_0(\hat{H}, \hat{\mathcal{B}}_{LS}(\hat{H})) \geq \det \mathrm{MCD}_q(\hat{\mathcal{B}}_{LS}(\hat{H})),$$

where we applied the definition of $\hat{H}$ and MCD_q. We conclude that $\hat{\mathcal{B}}_{LS}(\hat{H}) \in \underset{\mathcal{B}}{\operatorname{argmin}} \det \mathrm{MCD}_q(\mathcal{B})$.

On the other hand, take now $\tilde{\mathcal{B}} \in \underset{\mathcal{B}}{\operatorname{argmin}} \det \mathrm{MCD}_q(\mathcal{B})$ By definition of MCD, there exists a $\tilde{H} \in \mathcal{H}$ such that $\mathrm{MCD}_q(\tilde{\mathcal{B}}) = \mathrm{Cov}_0(\tilde{H}, \tilde{\mathcal{B}})$ and in particular $\det \mathrm{Cov}_0(\tilde{H}, \tilde{\mathcal{B}}) \leq \det \mathrm{Cov}_0(\tilde{H}, \hat{\mathcal{B}}_{LS}(\tilde{H}))$. But since (11.5) also holds for the pair $(\tilde{H}, \tilde{\mathcal{B}})$, the uniqueness part of Lemma 1 gives $\tilde{\mathcal{B}} = \hat{\mathcal{B}}_{LS}(\tilde{H})$. It then follows that for any other $H \in \mathcal{H}$ we have

$$\det \hat{\Sigma}_{LS}(H) = \det \mathrm{Cov}_0(H, \hat{\mathcal{B}}_{LS}(H)) \geq \det \mathrm{MCD}_q(\hat{\mathcal{B}}_{LS}(H)) \geq \det \mathrm{MCD}_q(\tilde{\mathcal{B}}) = \det \hat{\Sigma}_{LS}(\tilde{H}).$$

Hence, we have that $\tilde{H} \in \underset{H}{\operatorname{argmin}} \det \Sigma_{LS}(H)$ which ends the proof. □

Proof of Proposition 2: For any $H \in \mathcal{H}$ denote $\tilde{\Sigma}_{LS}(H) := (\det \hat{\Sigma}_{LS}(H))^{-1/q} \hat{\Sigma}_{LS}(H)$ such that $\det \tilde{\Sigma}_{LS}(H) = 1$. We first give the following equations which will be useful to prove the result. Using properties of traces, we find that

$$\begin{aligned} \frac{1}{h} \sum_{j \in H} d_j^2(\hat{\mathcal{B}}_{LS}(H), \hat{\Sigma}_{LS}(H)) &= \frac{1}{h} \mathrm{tr} \sum_{j \in H} \hat{\Sigma}_{LS}(H)^{-1} r_j(\hat{\mathcal{B}}_{LS}(H)) r_j(\hat{\mathcal{B}}_{LS}(H))^t \\ &= \mathrm{tr}\, \hat{\Sigma}_{LS}(H)^{-1} \hat{\Sigma}_{LS}(H) = q. \end{aligned} \tag{11.6}$$

We also have that

$$\sum_{j \in H} d_j^2(\hat{\mathcal{B}}_{LS}(H), \hat{\Sigma}_{LS}(H)) = (\det \hat{\Sigma}_{LS}(H))^{-1/q} \sum_{j \in H} d_j^2(\hat{\mathcal{B}}_{LS}(H), \tilde{\Sigma}_{LS}(H)) \tag{11.7}$$

Combining (11.7) with (11.6) yields

$$\sum_{j \in H} d_j^2(\hat{\mathcal{B}}_{LS}(H), \tilde{\Sigma}_{LS}(H)) = hq \det \hat{\Sigma}_{LS}(H))^{1/q}. \tag{11.8}$$

We first prove that for any $\hat{H} \in \underset{H}{\operatorname{argmin}} \det \hat{\Sigma}_{LS}(H)$ we have that $\hat{\mathcal{B}}_{LS}(\hat{H}) \in \{\tilde{\mathcal{B}} \,|\, (\tilde{\mathcal{B}}, \tilde{\Sigma}) \in \underset{\mathcal{B}, \Sigma; |\Sigma|=1}{\operatorname{argmin}} \sum_{j=1}^h d_{j:n}^2(\mathcal{B}, \Sigma)\}$. Take $\hat{H} \in \underset{H}{\operatorname{argmin}} \det \hat{\Sigma}_{LS}(H)$ and denote

$$H' := \{j |\, d_j(\hat{\mathcal{B}}_{LS}(\hat{H}), \tilde{\Sigma}_{LS}(\hat{H})) \leq d_{h:n}(\hat{\mathcal{B}}_{LS}(\hat{H}), \tilde{\Sigma}_{LS}(\hat{H}))\} \in \mathcal{H}.$$

the set of indices corresponding to the first h ordered squared distances of the residuals. Now suppose that

$$\sum_{j=1}^{h} d_{j:n}^2(\hat{\mathcal{B}}_{LS}(\hat{H}), \tilde{\Sigma}_{LS}(\hat{H})) = \sum_{j \in H'} d_j^2(\hat{\mathcal{B}}_{LS}(\hat{H}), \tilde{\Sigma}_{LS}(\hat{H})) < \sum_{j \in \hat{H}} d_j^2(\hat{\mathcal{B}}_{LS}(\hat{H}), \tilde{\Sigma}_{LS}(\hat{H})).$$

Using (11.7) and (11.8), this yields $\frac{1}{h} \sum_{j \in H'} d_j^2(\hat{\mathcal{B}}_{LS}(\hat{H}), \hat{\Sigma}_{LS}(\hat{H})) < q$. Therefore, there exists a constant $0 < c < 1$ such that $\frac{1}{h} \sum_{j \in H'} d_j^2(\hat{\mathcal{B}}_{LS}(\hat{H}), c\hat{\Sigma}_{LS}(\hat{H})) = q$. It then follows from Lemma 1 that $\det \hat{\Sigma}_{LS}(H') < \det c\hat{\Sigma}_{LS}(\hat{H}) < \det \hat{\Sigma}_{LS}(\hat{H})$ which is a contradiction, so we conclude that

$$\sum_{j=1}^{h} d^2_{j:n}(\hat{\mathcal{B}}_{LS}(\hat{H}), \tilde{\Sigma}_{LS}(\hat{H})) = \sum_{j \in \hat{H}} d^2_j(\hat{\mathcal{B}}_{LS}(\hat{H}), \tilde{\Sigma}_{LS}(\hat{H})). \tag{11.9}$$

Now suppose that there exists some $\mathcal{B} \in I\!R^{p \times q}$ and $\Sigma \in \text{PDS}(q)$ with $\det \Sigma = 1$ such that

$$\sum_{j=1}^{h} d^2_{j:n}(\mathcal{B}, \Sigma) < \sum_{j=1}^{h} d^2_{j:n}(\hat{\mathcal{B}}_{LS}(\hat{H}), \tilde{\Sigma}_{LS}(\hat{H})) \tag{11.10}$$

Denote $H_1 := \{j |\, d_j(\mathcal{B}, \Sigma) \leq d_{h:n}(\mathcal{B}, \Sigma)\} \in \mathcal{H}$ the set of indices corresponding to the first h ordered squared distances of the residuals and suppose that

$$\sum_{j=1}^{h} d^2_{j:n}(\mathcal{B}, \Sigma) = \sum_{j \in H_1} d^2_j(\mathcal{B}, \Sigma) < \sum_{j \in H_1} d^2_j(\hat{\mathcal{B}}_{LS}(H_1), \tilde{\Sigma}_{LS}(H_1)).$$

Using (11.8) this implies that $\frac{1}{h} \sum_{j \in H_1} d^2_j(\mathcal{B}, \det \hat{\Sigma}_{LS}(H_1)^{1/q} \Sigma) < q$. Hence, there exists a constant $0 < c < 1$ such that $\frac{1}{h} \sum_{j \in H_1} d^2_j(\mathcal{B}, c \det \hat{\Sigma}_{LS}(H_1)^{1/q} \Sigma) = q$. From Lemma 1 it follows that $\det \hat{\Sigma}_{LS}(H_1) < \det\, (c \det \hat{\Sigma}_{LS}(H_1)^{1/q} \Sigma) = c^q \det \hat{\Sigma}_{LS}(H_1)$ which is a contradiction, so we have that

$$\sum_{j=1}^{h} d^2_{j:n}(\mathcal{B}, \Sigma) \geq \sum_{j \in H_1} d^2_j(\hat{\mathcal{B}}_{LS}(H_1), \tilde{\Sigma}_{LS}(H_1)). \tag{11.11}$$

From (11.9) and (11.11) it follows that the inequality (11.10) implies that

$$\sum_{j \in H_1} d^2_j(\hat{\mathcal{B}}_{LS}(H_1), \tilde{\Sigma}_{LS}(H_1)) < \sum_{j \in \hat{H}} d^2_j(\hat{\mathcal{B}}_{LS}(\hat{H}), \tilde{\Sigma}_{LS}(\hat{H})). \tag{11.12}$$

But, using (11.8), this can be rewritten as $hq \det \hat{\Sigma}_{LS}(H_1)^{1/q} < hq \det \hat{\Sigma}_{LS}(\hat{H})^{1/q}$. Hence, we obtain $\det \hat{\Sigma}_{LS}(H_1) < \det \hat{\Sigma}_{LS}(\hat{H})$ which is a contradiction since $\hat{H} \in \underset{H}{\text{argmin}} \det \hat{\Sigma}_{LS}(H)$. Therefore, we conclude that

$$\sum_{j=1}^{h} d^2_{j:n}(\hat{\mathcal{B}}_{LS}(\hat{H}), \tilde{\Sigma}_{LS}(\hat{H})) \leq \sum_{j=1}^{h} d^2_{j:n}(\mathcal{B}, \Sigma)$$

for all $\mathcal{B} \in I\!R^{p \times q}$ and $\Sigma \in \text{PDS}(q)$ with $\det \Sigma = 1$ and thus we have $\hat{\mathcal{B}}_{LS}(\hat{H}) \in \{\tilde{\mathcal{B}} |(\tilde{\mathcal{B}}, \tilde{\Sigma}) \in \underset{\mathcal{B}, \Sigma; |\Sigma|=1}{\text{argmin}} \sum_{j=1}^{h} d^2_{j:n}(\mathcal{B}, \Sigma)\}$.

We now prove that for any $(\tilde{\mathcal{B}}, \tilde{\Sigma}) \in \underset{\mathcal{B}, \Sigma; |\Sigma|=1}{\text{argmin}} \sum_{j=1}^{h} d^2_{j:n}(\mathcal{B}, \Sigma)$ there exists a $\tilde{H} \in \mathcal{H}$ such that $\tilde{\mathcal{B}} = \hat{\mathcal{B}}_{LS}(\tilde{H})$ and $\tilde{H} \in \underset{H}{\text{argmin}} \det \hat{\Sigma}_{LS}(H)$. Denote $\tilde{H} := \{j |\, d_j(\tilde{\mathcal{B}}, \tilde{\Sigma}) \leq d_{h:n}(\tilde{\mathcal{B}}, \tilde{\Sigma})\} \in \mathcal{H}$ the set of indices corresponding to the first h ordered squared distances of the residuals, then we have that

$$\sum_{j=1}^{h} d^2_{j:n}(\tilde{\mathcal{B}}, \tilde{\Sigma}) = \sum_{j \in \tilde{H}} d^2_j(\tilde{\mathcal{B}}, \tilde{\Sigma}) \leq \sum_{j \in \tilde{H}} d^2_j(\hat{\mathcal{B}}_{LS}(\tilde{H}), \tilde{\Sigma}_{LS}(\tilde{H})). \tag{11.13}$$

Using (11.8) it follows that $\frac{1}{h}\sum_{j\in\tilde{H}} d_j^2(\tilde{\mathcal{B}}, \det\hat{\Sigma}_{LS}(\tilde{H})^{1/q}\,\tilde{\Sigma}) \leq q$. Hence, there exists a constant $0 < c \leq 1$ such that $\frac{1}{h}\sum_{j\in\tilde{H}} d_j^2(\tilde{\mathcal{B}}, c\det\hat{\Sigma}_{LS}(\tilde{H})^{1/q}\,\tilde{\Sigma}) = q$. From Lemma 1 we then obtain that $\det\hat{\Sigma}_{LS}(\tilde{H}) \leq \det(c\det\hat{\Sigma}_{LS}(\tilde{H})^{1/q}\,\tilde{\Sigma}) = c^q\det\hat{\Sigma}_{LS}(\tilde{H})$ which is a contradiction unless if $c = 1$ and by Lemma 1 (uniqueness) we then have that $\tilde{\mathcal{B}} = \hat{\mathcal{B}}_{LS}(\tilde{H})$ and $\tilde{\Sigma} = \tilde{\Sigma}_{LS}(\tilde{H})$. For any $H \in \mathcal{H}$ we now have that

$$\sum_{j=1}^{h} d_{j:n}^2(\tilde{\mathcal{B}}, \tilde{\Sigma}) = \sum_{j\in\tilde{H}} d_j^2(\hat{\mathcal{B}}_{LS}(\tilde{H}), \tilde{\Sigma}_{LS}(\tilde{H})) \leq \sum_{j\in H} d_j^2(\hat{\mathcal{B}}_{LS}(H), \tilde{\Sigma}_{LS}(H))$$

By using (11.8) the inequality can be rewritten as $hq\det\hat{\Sigma}_{LS}(\tilde{H})^{1/q} \leq hq\det\hat{\Sigma}_{LS}(H)^{1/q}$ which yields $\det\hat{\Sigma}_{LS}(\tilde{H}) \leq \det\hat{\Sigma}_{LS}(H)$ for all $H \in \mathcal{H}$. Therefore, we conclude that $\tilde{H} \in \underset{H}{\operatorname{argmin}} \det\hat{\Sigma}_{LS}(H)$ which ends the proof. □

Proof of Theorem 1: We first prove that $\varepsilon_n^*(\hat{\mathcal{B}}_{MLTS}, Z_n) \geq \min(n-h+1, h-k(Z_n))/n$. We will show that there exists a value $\bar{M}$, which only depends on Z_n, such that for every Z_n' obtained by replacing at most $m = \min(n-h+1, h-k(Z_n)) - 1$ observations from Z_n we have that $\|\hat{\mathcal{B}}_{MLTS}(Z_n')\| \leq \bar{M}$. The matrix norm we use here is $\|A\| = \sup\limits_{\|u\|=1} \|Au\|$ where $u \in I\!R^q$ and $A \in I\!R^{p\times q}$. Sometimes we will also use the L_2-norm $\|A\|_2 = (\sum_{i,j}|a_{ij}|^2)^{1/2}$. Since all norms on $I\!R^{p\times q}$ are topologically equivalent there exist values $\alpha_1, \alpha_2 > 0$ such that $\alpha_1\|A\| \leq \|A\|_2 \leq \alpha_2\|A\|$ for all $A \in I\!R^{p\times q}$.

Let J be a subset of size $k(Z_n)+1$. Then there cannot be a hyperplane such that all x_j with $j \in J$ are on it. Therefore

$$c_1(J) = \frac{1}{2}\inf_{\|\gamma\|=1}\max_{j\in J}|\gamma^t x_j| > 0$$

where $\gamma \in I\!R^p$. Furthermore it is excluded that there exists a $\mathcal{B} \in I\!R^{p+q}$ such that $y_j - \mathcal{B}^t x_j$ for all $j \in J$ are lying on a $(q-1)$ dimensional hyperplane. Indeed, otherwise there exists an $\alpha \in I\!R^q$ such that for all $j \in J$ we have $\alpha^t(y_j - \mathcal{B}^t x_j) = \alpha^t y_j - \gamma^t x_j = 0$ where $\gamma = \mathcal{B}\alpha$. However, this contradicts the assumption $\#J = k(Z_n)+1$. Since for all $\mathcal{B} \in I\!R^{p+q}$ the $r_j := y_j - \mathcal{B}^t x_j$ are not lying on a $(q-1)$ dimensional hyperplane, we have that

$$c_2(J) = \inf_{\mathcal{B}\in I\!R^{p+q}} \lambda_{\min}\operatorname{Cov}_0(\{r_j; j \in J\}) > 0$$

where $\operatorname{Cov}_0(\{r_j; j \in J\}) = \frac{1}{k(Z_n)+1}\sum_{j\in J} r_j r_j^t$ and $\lambda_{\min}$ denotes the smallest eigenvalue of that matrix. Denote

$$c = \min_J\,(\min(c_1(J), c_2(J))) > 0 \qquad (11.14)$$

where the minimum is over all subsets J of size $k(Z_n)+1$ and define

$$M = \sup_{H \in \mathcal{H}} \|\mathcal{B}_{LS}(H)^t\| < \infty \tag{11.15}$$

since no h points of $\{x_i; i = 1, \ldots, n\}$ are lying on the same hyperplane ($k(Z_n) < h$). Let $N_y = \max_{1 \leq i \leq n} \|y_i\|$ and $N_x = \max_{1 \leq i \leq n} \|x_i\|$. Put $V = (N_y + M\, N_x)^{2q}$ and

$$\bar{M} = ((V\, h\, (\frac{h}{k(Z_n)+1}\, c)^{1-q})^{1/2} + N_y)\frac{1}{\alpha_1 c} \tag{11.16}$$

Now take a dataset Z'_n obtained by replacing m observations from Z_n and suppose $\|\hat{\mathcal{B}}_{MLTS}(Z'_n)\| > \bar{M}$. First of all, there exists a subset $H_1 \in \mathcal{H}$ containing indices only corresponding to data points of the original dataset Z_n. Using lemma 5.1 of Lopuhaä and Rousseeuw (1991, page 244) and properties of norms it follows that

$$\begin{aligned}
\det(\hat{\Sigma}_{LS}(H_1)) &\leq \lambda_{\max}(\mathrm{cov}(\{r_j(\hat{\mathcal{B}}_{LS}(H_1)); j \in H_1\})^q \\
&\leq (\frac{1}{h}\sum_{j \in H_1} \lambda_{\max}(r_j(\hat{\mathcal{B}}_{LS}(H_1))r_j(\hat{\mathcal{B}}_{LS}(H_1))^t))^q \\
&= (\frac{1}{h}\sum_{j \in H_1} \|r_j(\hat{\mathcal{B}}_{LS}(H_1))\|^2)^q \\
&\leq (\frac{1}{h}\sum_{j \in H_1} (\|y_j\| + \|\hat{\mathcal{B}}_{LS}(H_1)^t x_j\|)^2)^q \\
&\leq (N_y + M\, N_x)^{2q} \\
&= V
\end{aligned} \tag{11.17}$$

where $\lambda_{\max}$ denotes the largest eigenvalue of a matrix. Now let H_2 be the optimal subset corresponding to $\hat{\mathcal{B}}_{MLTS}(Z'_n)$ such that $\hat{\mathcal{B}}_{MLTS}(Z'_n) = \hat{\mathcal{B}}_{LS}(H_2) := \mathcal{B}_2$. Since $h - m \geq k(Z_n) + 1$ the set H_2 contains a subset $\bar{J}$ of size $k(Z_n)+1$ corresponding to original observations of Z_n. Using lemma 5.1 of Lopuhaä and Rousseeuw (1991, page 244) we obtain

$$\begin{aligned}
\lambda_{\min}(\hat{\Sigma}_{LS}(H_2)) &= \lambda_{\min}(\mathrm{cov}(\{y_j - \mathcal{B}_2^t x_j; j \in H_2\})) \\
&\geq \frac{k(Z_n)+1}{h}\lambda_{\min}(\mathrm{Cov}_0(\{y_j - \mathcal{B}_2^t x_j; j \in \bar{J}\})) \\
&\geq \frac{k(Z_n)+1}{h} c_2(\bar{J}) \\
&\geq \frac{k(Z_n)+1}{h} c
\end{aligned} \tag{11.18}$$

On the other hand,

$$\lambda_{\max}(\hat{\Sigma}_{LS}(H_2)) = \sup_{\|u\|=1} \frac{1}{h}\sum_{j \in H_2} u^t(y_j - \mathcal{B}_2^t x_j)(y_j - \mathcal{B}_2^t x_j)^t u. \tag{11.19}$$

By definition of $c_1(\bar{J})$ there exists at least one index $j_0 \in \bar{J} \subset H_2$ such that

$$\begin{aligned}
\|\mathcal{B}_2^t x_{j0}\|^2 &= \sum_{j=1}^{q} |\mathcal{B}_{2j} x_{j0}|^2 \\
&\geq \sum_{j=1}^{q} \|\mathcal{B}_{2j}\|^2 \, c_1(\bar{J})^2 \\
&= (\|\mathcal{B}_2\|_2 \, c_1(\bar{J}))^2 \\
&\geq (\alpha_1 \|\mathcal{B}_2\| c_1(\bar{J}))^2
\end{aligned}$$

which yields $\|\mathcal{B}_2^t x_{j_0}\| > \alpha_1 \bar{M} c$. Since by definition $\alpha_1 \bar{M} c \geq N_y$ we obtain $\|y_{j_0} - \mathcal{B}_2^t x_{j_0}\| \geq |\, \|y_{j_0}\| - \|\mathcal{B}_2^t x_{j_0}\| \,| > \alpha_1 \bar{M} c - N_y$. By taking $u = \frac{y_{j_0} - \mathcal{B}_2^t x_{j_0}}{\|y_{j_0} - \mathcal{B}_2^t x_{j_0}\|}$ it follows from (11.19) that

$$\lambda_{\max}(\hat{\Sigma}_{LS}(H_2)) \geq \|y_{j_0} - \mathcal{B}_2^t x_{j_0}\|^2/h > (\alpha_1 \bar{M} c - N_y)^2/h. \tag{11.20}$$

Combining (11.20) and (11.18) yields

$$\det(\hat{\Sigma}_{LS}(H_2)) > \frac{1}{h}(\alpha_1 \bar{M} c - N_y)^2 (\frac{k(Z_n)+1}{h} c)^{q-1} = V$$

by definition of $\bar{M}$. Together with (11.17) this implies $\det(\hat{\Sigma}_{LS}(H_2)) > \det(\hat{\Sigma}_{LS}(H_1))$ which contradicts the definition of $\hat{\mathcal{B}}_{MLTS}(Z_n')$, so we conclude that $\|\hat{\mathcal{B}}_{MLTS}(Z_n')\| \leq \bar{M}$.

We now prove that also $\varepsilon_n^*(\hat{\mathcal{B}}_{MLTS}, Z_n) \leq \min(n-h+1, h-k(Z_n))/n$. First we show that $\varepsilon_n^*(\hat{\mathcal{B}}_{MLTS}, Z_n) \leq (n-h+1)/n$. Indeed, if we replace $n-h+1$ points of Z_n then the optimal subset H_2 of Z_n' will contain at least one outlier and we know that least squares can explode in the presence of even a single outlier. It then follows that also $\hat{\mathcal{B}}_{MLTS}(Z_n')$ explodes.

Now we show that $\varepsilon_n^*(\hat{\mathcal{B}}_{MLTS}, Z_n) \leq (h-k(Z_n))/n$. Denote $\tilde{J} \subset \{1, \ldots, n\}$ the set of indices corresponding to the $k(Z_n)$ observations from Z_n lying on a hyperplane of $\mathbb{R}^{p+q}$. Then there exist a $\alpha \in \mathbb{R}^q$ and $\gamma \in \mathbb{R}^p$ such that $\alpha^t y_j - \gamma^t x_j = 0$ for all $j \in \tilde{J}$.

If $\alpha \neq 0$ then there exists a $\mathcal{B} \in \mathbb{R}^{p+q}$ such that $\mathcal{B}\alpha = \gamma$ which implies $\alpha^t(y_j - \mathcal{B}^t x_j) = 0$ for $j \in \tilde{J}$. Therefore, for $j \in \tilde{J}$ we have that $y_j - \mathcal{B}^t x_j \in S$ where S is a $(q-1)$ dimensional subspace of $\mathbb{R}^q$. Now take a $\mathcal{D} \in \mathbb{R}^{p \times q}$ with $\|\mathcal{D}\| = 1$ such that $\{\mathcal{D}^t x; x \in \mathbb{R}^p\} \subset S$. Now replace $m = h - k(Z_n)$ observations of Z_n, not lying on S, by $(x_0, (\mathcal{B} + \lambda\mathcal{D})^t x_0)$ for some arbitrarily chosen $x_0 \in \mathbb{R}^p$ and $\lambda \in \mathbb{R}$. Denote J_o the set of indices corresponding to the outliers. It follows that for the m outliers $r_j(\mathcal{B} + \lambda\mathcal{D}) = 0$ and for the $k(Z_n)$ points on S we have that $r_j(\mathcal{B} + \lambda\mathcal{D}) = y_j - \mathcal{B}^t x_j - \lambda\mathcal{D}^t x_j \in S$. Therefore $\{r_j(\mathcal{B} + \lambda\mathcal{D}); j \in \tilde{J} \cup J_o\}$ belongs to the subspace S, giving a zero determinant for the matrix $\text{cov}_0(\{r_j(\mathcal{B} + \lambda\mathcal{D}); j \in \tilde{J} \cup J_o\})$ Therefore, using Proposition 1 it follows that $\hat{\mathcal{B}}_{MLTS}(Z_n') = \mathcal{B} + \lambda\mathcal{D}$ which tends to infinity when $\lambda \to \infty$.

If $\alpha = 0$ then we have that $\gamma^t x_j = 0$ for all $j \in \tilde{J}$. Now replace $m = h - k(Z_n)$ other observations of Z_n by observations on the hyperplane $\gamma^t x = 0$. Denote H_2 the set of indices corresponding with observations of Z_n' such that $\gamma^t x = 0$. Since all these observations belong to a hyperplane of $I\!R^{p+q}$ we have that $\det \operatorname{cov}(\{y_j - \hat{\mathcal{B}}_{LS}(H_2)^t x_j; j \in H_2\}) = 0$. But since $\gamma^t x = 0$ is a vertical hyperplane we have $\|\hat{\mathcal{B}}_{LS}(H_2)\| = \infty$ and it follows that $\|\hat{\mathcal{B}}_{MLTS}(Z_n')\| = \infty$. □

Proof of Corollary 1. Since for $q = 1$ we have $\det(\hat{\Sigma}_{LS}(H_2)) = \lambda_{\max}(\hat{\Sigma}_{LS}(H_2))$, we do not need to establish the lower bound (11.18) and thus we do not need $c_2(\bar{J}) > 0$. To obtain $c_1(\bar{J}) > 0$ it suffices to consider datasets of size $k'(Z_n)+1$. Therefore, the result immediately follows from the previous proof if we replace $k(Z_n)$ by $k'(Z_n)$. □

Proof of Theorem 2. Using properties of traces we obtain

$$\frac{1}{h}\sum_{j\in H_2} d_j^2(\hat{\mathcal{B}}_2, \hat{\Sigma}_2) = \frac{1}{h}\mathrm{tr}\sum_{j\in H_2} r_j(\hat{\mathcal{B}}_2)^t \hat{\Sigma}_2^{-1} r_j(\hat{\mathcal{B}}_2) = \mathrm{tr}\,\hat{\Sigma}_2^{-1}\hat{\Sigma}_2 = \mathrm{tr}\, I_q = q \tag{11.21}$$

and similarly $\frac{1}{h}\sum_{j\in H_1} d_j^2(\hat{\mathcal{B}}_1, \hat{\Sigma}_1) = q$. By definition of H_2 we have

$$c := \frac{1}{hq}\sum_{j\in H_2} d_j^2(\hat{\mathcal{B}}_1, \hat{\Sigma}_1) \leq \frac{1}{hq}\sum_{j\in H_1} d_j^2(\hat{\mathcal{B}}_1, \hat{\Sigma}_1) = 1, \tag{11.22}$$

and also $c > 0$ since $\det(\hat{\Sigma}_2) > 0$. Combining (11.21) and (11.22) yields

$$\frac{1}{h}\sum_{j\in H_2} r_j(\hat{\mathcal{B}}_1)^t (c\hat{\Sigma}_1)^{-1} r_j(\hat{\mathcal{B}}_1) = \frac{1}{ch}\sum_{j\in H_2} d_j^2(\hat{\mathcal{B}}_1, \hat{\Sigma}_1) = \frac{cq}{c} = q. \tag{11.23}$$

From Lemma 1 it follows that $\det(\hat{\Sigma}_2) \leq \det(c\hat{\Sigma}_1)$ and (11.22) implies $\det(c\hat{\Sigma}_1) \leq \det(\hat{\Sigma}_1)$, hence $\det(\hat{\Sigma}_2) \leq \det(\hat{\Sigma}_1)$. Moreover, from Lemma 1 we know that $\det(\hat{\Sigma}_2) = \det(c\hat{\Sigma}_1)$ iff $\hat{\mathcal{B}}_2 = \hat{\mathcal{B}}_1$ and $\hat{\Sigma}_2 = c\hat{\Sigma}_1$. Furthermore, $\det(c\hat{\Sigma}_1) = \det(\hat{\Sigma}_1)$ iff $c = 1$. Therefore, $\det(\hat{\Sigma}_2) = \det(\hat{\Sigma}_1)$ iff $\hat{\mathcal{B}}_2 = \hat{\mathcal{B}}_1$ and $\hat{\Sigma}_2 = \hat{\Sigma}_1$. □

Proof of Lemma 2. Clearly, we have that $\mathcal{E} \in D_H(\alpha)$. Note that

$$\frac{1}{1-\alpha}\int_{\hat{A}} d^2(x,y)\,dH = \frac{1}{1-\alpha}\mathrm{tr}\int_{\hat{A}} d^2(x,y)\,dH = \mathrm{tr}\,(\Sigma_{\hat{A}}(H)^{-1}\Sigma_{\hat{A}}(H)) = \mathrm{tr}\, I_q = q$$

On the other hand, we have that

$$
\begin{aligned}
\int_{\mathcal{E}} d^2(x,y)\,dH &= \int_{\mathcal{E}\cap\hat{A}} d^2(x,y)\,dH + \int_{\mathcal{E}\setminus\hat{A}} d^2(x,y)\,dH \\
&\leq \int_{\mathcal{E}\cap\hat{A}} d^2(x,y)\,dH + D^2_\alpha P_H(\mathcal{E}\setminus\hat{A}) \\
&= \int_{\mathcal{E}\cap\hat{A}} d^2(x,y)\,dH + D^2_\alpha P_H(\hat{A}\setminus\mathcal{E}) \\
&\leq \int_{\mathcal{E}\cap\hat{A}} d^2(x,y)\,dH + \int_{\hat{A}\setminus\mathcal{E}} d^2(x,y)\,dH \\
&= \int_{\hat{A}} d^2(x,y)\,dH
\end{aligned}
$$

Therefore, there exists a $0 < c \leq 1$ such that

$$
\frac{1}{1-\alpha}\int_{\mathcal{E}}(y-(\mathcal{B}_{\hat{A}}(H))^t x)^t(c\,\Sigma_{\hat{A}}(H))^{-1}(y-(\mathcal{B}_{\hat{A}}(H))^t x)\,dH = q \tag{11.24}
$$

Since $\hat{A}$ is an MCD solution, we have that $\det(c\,\Sigma_{\hat{A}}(H)) \leq \det\Sigma_{\hat{A}}(H) \leq \det\Sigma_{\mathcal{E}}(H)$ which in combination with (11.24) contradicts lemma 1 unless if $\mathcal{B}_{\hat{A}}(H) = \mathcal{B}_{\mathcal{E}}(H)$ and $c\,\Sigma_{\hat{A}}(H) = \Sigma_{\mathcal{E}}(H)$. Then c should also be equal to 1. □

Proof of Theorem 3. First of all, due to equivariance, we may assume that $\mathcal{B} = 0$ and $\Sigma = I_q$, so $y = \varepsilon \sim F$. It now suffices to show that $\mathcal{B}_{LTS}(H) = 0$. Then we will have that $\Sigma_{LTS}(H)$ is the MCD functional at the distribution of $y - \mathcal{B}_{LTS}(H)^t x = y = \varepsilon$. Since the factor c_α makes the MCD Fisher-consistent at elliptical distributions (see Butler et al. 1993, Croux and Haesbroeck 1999) it will follow that $\Sigma_{LTS}(H) = I_q$. Lemma 2 shows that $\mathcal{B}_{LTS}$ is the least squares fit based solely on the cylinder $\mathcal{C} = \{(x,y) \in I\!R^{p+q}; (y - \mathcal{B}^t_{LTS}x)^t\Sigma^{-1}_{LTS}(y - \mathcal{B}^t_{LTS}x) \leq D^2_\alpha\}$. Therefore,

$$
\int_{\mathcal{C}} x(y-\mathcal{B}^t_{LTS}x)^t\,dH(x,y) = 0 \tag{11.25}
$$

Now suppose that $\mathcal{B}_{LTS} \neq 0$. Let $\lambda_1,\ldots,\lambda_q$ be the eigenvalues of Σ_{LTS} and $v_1,\ldots,v_q$ the corresponding eigenvectors. There will be at least one $1 \leq j \leq q$ such that $\mathcal{B}_{LTS}v_j \neq 0$. (Note that $\mathcal{B}_{LTS}$ is not necessarily of full rank.) Fix this j. From (11.25) it follows that we should have

$$
\int_{\mathcal{E}} v^t_j(\mathcal{B}^t_{LTS}x)(y-\mathcal{B}^t_{LTS}x)^t v_j\,dF(y)\,dG(x) = 0
$$

which can be rewritten as

$$
\int_{I\!R^p} v^t_j(\mathcal{B}^t_{LTS}x)I(x)\,dG(x) = 0 \tag{11.26}
$$

with

$$
I(x) = \int_{\mathcal{C}_x}(y-\mathcal{B}^t_{LTS}x)^t v_j\,dF(y),
$$

where $\mathcal{C}_x = \{y \in I\!R^q | (x,y) \in \mathcal{C}\}$. Fix x and set $d = (d_1, \ldots, d_q)^t := \mathcal{B}_{LTS}^t x$. Since y is spherically symmetrically distributed, for computing $I(x)$ we may assume w.l.o.g. that $\Sigma_{LTS} = \text{diag}(\lambda_1, \ldots, \lambda_q)$ as well as $v_j = (1, 0, \ldots, 0)$. For every $d_1 - \sqrt{c\lambda_1} \leq y_1 \leq d_1 + \sqrt{c\lambda_1}$ denote

$$\mathcal{C}(y_1) = \left\{ (y_2, \ldots, y_q) \in I\!R^{q-1} | \sum_{j=2}^{q} \frac{(y_j - d_j)^2}{\lambda_j} \leq c - \frac{(y_1 - d_1)^2}{\lambda_1} \right\}$$

where $c := D_\alpha^2 > 0$. Then we can rewrite $I(x)$ as

$$\begin{aligned} I(x) &= \int_{d_1 - \sqrt{c\lambda_1}}^{d_1 + \sqrt{c\lambda_1}} \int_{\mathcal{C}(y_1)} (y_1 - d_1) g(y_1^2 + \cdots + y_q^2)\, dy_1 \ldots dy_q \\ &= \int_{-\sqrt{c\lambda_1}}^{\sqrt{c\lambda_1}} t \int_{\mathcal{C}(d_1+t)} g((d_1+t)^2 + y_2^2 + \cdots + y_q^2)\, dy_2 \ldots dy_q\, dt. \end{aligned}$$

Since $\mathcal{C}(d_1 + t) = \mathcal{C}(d_1 - t)$ it follows that

$$I(x) = \int_0^{\sqrt{c\lambda_1}} t \int_{\mathcal{C}(d_1+t)} g\left((d_1+t)^2 + y_2^2 + \cdots + y_q^2\right) - g\left((d_1-t)^2 + y_2^2 + \cdots + y_q^2\right) dy_2 \ldots dy_q\, dt.$$

If $d_1 > 0$ we have $(d_1+t)^2 + y_2^2 + \cdots + y_q^2 > (d_1-t)^2 + y_2^2 + \cdots + y_q^2$ (for $t > 0$) and since g is strictly decreasing this implies $I(x) < 0$. Similarly, we can show that $d_1 < 0$ implies $I(x) > 0$ and that $d_1 = 0$ yields $I(x) = 0$. Hence, we have shown that $v_j^t(\mathcal{B}_{LTS}^t x) > 0$ implies $I(x) < 0$ and if $v_j^t(\mathcal{B}_{LTS}^t x) = 0$, then $I(x) > 0$. Also, $v_j^t(\mathcal{B})_{LTS}^t x) = 0$ implies $I(x) = 0$. However, due to condition (5.6), the latter event occurs with probability less than $1 - \alpha$. Therefore, we obtain $\int_{I\!R^p} v_j^t \mathcal{B}_{LTS}^t x\, I(x)\, dG(x) < 0$ which contradicts (11.26), so we conclude that $\mathcal{B}_{LTS} = 0$. □

Proof of Theorem 4. Consider the contaminated distribution $H_\varepsilon = (1-\varepsilon)H_0 + \varepsilon\Delta_{z_0}$ with $z_0 = (x_0, y_0)$ and denote $\mathcal{B}_\varepsilon := \mathcal{B}_{LTS}(H_\varepsilon)$ and $\Sigma_\varepsilon := \Sigma_{LTS}(H_\varepsilon)$. Then (5.3) results in

$$\hat{\mathcal{B}}_\varepsilon = \left(\int_{\hat{A}_\varepsilon} x x^t\, dH_\varepsilon(x,y) \right)^{-1} \int_{\hat{A}_\varepsilon} x y^t\, dH_\varepsilon(x,y)$$

where $\hat{A}_\varepsilon \in \mathcal{D}_{H_\varepsilon}(\alpha)$ is an MLTS solution. Differentiating w.r.t. ε and evaluating at 0 yields

$$IF(z_0; \mathcal{B}_{LTS}, H_0) = \left(\int_{\hat{A}} x x^t\, dH_0(z) \right)^{-1} \frac{\partial}{\partial\varepsilon} \int_{\hat{A}_\varepsilon} x y^t\, dH_\varepsilon(z)_{|_{\varepsilon=0}} + \frac{\partial}{\partial\varepsilon} \left[\left(\int_{\hat{A}_\varepsilon} x x^t\, dH_\varepsilon(z) \right)^{-1} \right]_{|_{\varepsilon=0}} \int_{\hat{A}} x y^t\, dH_0(z)$$

Lemma 2 combined with Fisher-consistency yields that $\hat{A} = \{(x,y) \in I\!R^{p+q}; y^t y \leq q_\alpha\}$ where $q_\alpha = (D_F^2)^{-1}(1-\alpha)$ with $D_F^2(t) = P_F(\|y\|^2 \leq t)$. Hence $\hat{A} = I\!R^p \times \{y \in I\!R^q; \|y\|^2 \leq q_\alpha\} =: I\!R^p \times A$. This implies

$$\int_{\hat{A}} x y^t\, dH_0(z) = \int_{I\!R^p} x\, dG(x) \int_A y^t\, dF(y) = 0$$

by symmetry of F and

$$\int_{\hat{A}} xx^t \, dH_0(z) = \int_{I\!R^p} xx^t \, dG(x) \int_A dF(y) = E_G[xx^t]\,(1-\alpha)$$

Therefore, we obtain

$$\begin{aligned} IF(z_0; \mathcal{B}_{LTS}, H_0) &= \frac{E_G[xx^t]^{-1}}{1-\alpha} \frac{\partial}{\partial \varepsilon} \int_{\hat{A}_\varepsilon} xy^t \, dH_\varepsilon(z)_{|_{\varepsilon=0}} \\ &= \frac{E_G[xx^t]^{-1}}{1-\alpha} \frac{\partial}{\partial \varepsilon} \left((1-\varepsilon) \int_{\hat{A}_\varepsilon} xy^t \, dH_0(z) + \varepsilon x_0 y_0^t I(z_0 \in \hat{A}_\varepsilon) \right)_{|_{\varepsilon=0}} \\ &= \frac{E_G[xx^t]^{-1}}{1-\alpha} \left(x_0 y_0^t I(\|y_0\|^2 \leq q_\alpha) + \frac{\partial}{\partial \varepsilon} \int_{\hat{A}_\varepsilon} xy^t \, dH_0(z) \right). \qquad (11.27) \end{aligned}$$

Similarly to Proposition 1 of Croux and Haesbroeck (1999), it can be shown that Lemma 2 still holds for contaminated distributions H_ε. Let us denote $d_\varepsilon^2(x,y) = (y - \mathcal{B}_\varepsilon^t x)^t \Sigma_\varepsilon^{-1} (y - \mathcal{B}_\varepsilon^t x)$, then it follows that $\hat{A}_\varepsilon = \{(x,y) \in I\!R^{p+q}; d_\varepsilon^2(x,y) \leq q_\alpha(\varepsilon)\}$ where $q_\alpha(\varepsilon) = (D_{H_\varepsilon}^2)^{-1}(1-\alpha)$ with $D_{H_\varepsilon}^2(t) = P_{H_\varepsilon}(d_\varepsilon^2(x,y) \leq t)$. For x fixed we define the ellipsoid $\mathcal{E}_{\varepsilon,x} := \{y \in I\!R^q; d_\varepsilon^2(x,y) \leq q_\alpha(\varepsilon)\}$. Then it follows that

$$\int_{\hat{A}_\varepsilon} xy^t \, dH_0(z) = \int_{I\!R^p} \int_{\mathcal{E}_{\varepsilon,x}} xy^t \, dF(y) dG(x) = \int_{I\!R^p} x \left(\int_{\mathcal{E}_{\varepsilon,x}} y\, g(y^t y) \, dy \right)^t dG(x). \qquad (11.28)$$

Using the transformation $v = \Sigma_\varepsilon^{-1/2}(y - \mathcal{B}_\varepsilon^t x)$, we obtain that

$$I(\varepsilon) := \int_{\mathcal{E}_{\varepsilon,x}} y\, g(y^t y)\, dy = \det(\Sigma_\varepsilon)^{1/2} \int_{\|v\|^2 \leq q_\alpha(\varepsilon)} (\Sigma_\varepsilon^{1/2} v + \mathcal{B}_\varepsilon^t x) g((\Sigma_\varepsilon^{1/2} v + \mathcal{B}_\varepsilon^t x)^t (\Sigma_\varepsilon^{1/2} v + \mathcal{B}_\varepsilon^t x))\, dv.$$

Rewriting this expression in polar coordinates $v = r\, e(\theta)$ where $r \in [0, \sqrt{q_\alpha(\varepsilon)}]$, $e(\theta) \in S^{q-1}$ and $\theta = (\theta_1, \ldots, \theta_{q-1}) \in \Theta = [0, \pi[\times \cdots \times [0, \pi[\times [0, 2\pi[$, yields

$$I(\varepsilon) = \det(\Sigma_\varepsilon)^{1/2} \int_0^{\sqrt{q_\alpha(\varepsilon)}} \int_\Theta J(\theta, r)(r\Sigma_\varepsilon^{1/2} e(\theta) + \mathcal{B}_\varepsilon^t x) g((r\Sigma_\varepsilon^{1/2} e(\theta) + \mathcal{B}_\varepsilon^t x)^t (r\Sigma_\varepsilon^{1/2} e(\theta) + \mathcal{B}_\varepsilon^t x))\, dr d\theta,$$

where $J(\theta, r)$ is the Jacobian of the transformation into polar coordinates. Applying Leibniz' formula to this expression and using the symmetry of F results in

$$\frac{\partial}{\partial \varepsilon} I(\varepsilon)_{|_{\varepsilon=0}} = \int_{\|v\|^2 \leq q_\alpha} \frac{\partial}{\partial \varepsilon} \left((\Sigma_\varepsilon^{1/2} v + \mathcal{B}_\varepsilon^t x) g((\Sigma_\varepsilon^{1/2} v + \mathcal{B}_\varepsilon^t x)^t (\Sigma_\varepsilon^{1/2} v + \mathcal{B}_\varepsilon^t x)) \right)_{|_{\varepsilon=0}} dv \qquad (11.29)$$

The derivative on the right hand side becomes

$$\frac{\partial}{\partial \varepsilon} \{ (\Sigma_\varepsilon^{1/2} v + \mathcal{B}_\varepsilon^t x) g((\Sigma_\varepsilon^{1/2} v + \mathcal{B}_\varepsilon^t x)^t (\Sigma_\varepsilon^{1/2} v + \mathcal{B}_\varepsilon^t x)) \}_{|_{\varepsilon=0}} =$$

$$\{ IF(z_0; \Sigma_{LTS}^{1/2}, H_0) v + IF(z_0; \mathcal{B}_{LTS}, H_0)^t x \} g(v^t v) + 2\, v g'(v^t v) \{ (v^t IF(z_0; \Sigma_{LTS}^{1/2}, H_0) v + v^t IF(z_0; \mathcal{B}_{LTS}, H_0)^t x \} \qquad (11.30)$$

Since $\int_{\|v\|^2\leq q_\alpha} vg(v^tv)\,dv$ and $\int_{\|v\|^2\leq q_\alpha} vg'(v^tv)v^tIF(z_0;\Sigma_{LTS}^{1/2},H_0)v\,dv$ are zero due to symmetry of F, the terms in (11.30) including $IF(z_0;\Sigma_{LTS}^{1/2},H_0)$ give a zero contribution to the integral in (11.29). It follows that

$$\begin{aligned}\frac{\partial}{\partial\varepsilon}I(\varepsilon)_{|_{\varepsilon=0}} &= (1-\alpha)IF(z_0;\mathcal{B}_{LTS},H_0)^tx + 2\int_{\|v\|^2\leq q_\alpha} g'(v^tv)vv^t\,dv\,IF(z_0;\mathcal{B}_{LTS},H_0)^tx \\ &= [(1-\alpha)+2c_2]\,IF(z_0;\mathcal{B}_{LTS},H_0)^tx\end{aligned}$$

where $c_2 = \int_{\|v\|^2\leq q_\alpha} g'(v^tv)v_1^2\,dv$ can be rewritten in the form given in Theorem 4 by using polar coordinates. From (11.28) we now obtain that

$$\frac{\partial}{\partial\varepsilon}\int_{\hat{A}_\varepsilon} xy^t\,dH_0(z)_{|_{\varepsilon=0}} = [(1-\alpha)+2c_2]\,E_G[xx^t]IF(z_0;\mathcal{B}_{LTS},H_0). \qquad (11.31)$$

Substituting (11.31) in (11.27) yields

$$(1-\alpha)IF(z_0;\mathcal{B}_{LTS},H_0) = E_G[xx^t]^{-1}xy^tI(\|y\|^2\leq q_\alpha) + [(1-\alpha)+2c_2]\,IF(z_0;\mathcal{B})LTS,H_0)$$

which results in

$$IF(z_0;\mathcal{B}_{LTS},H_0) = E_G[xx^t]^{-1}\frac{xy^t}{-2c_2}I(\|y\|^2\leq q_\alpha) \qquad \square$$

References

Bai, Z.D., Chen, N.R., Miao, B.Q., and Rao, C.R. (1990), "Asymptotic Theory of Least Distance Estimate in Multivariate Linear Models," *Statistics*, 21, 503-519.

Bilodeau, M. and Duchesne P. (2000), "Robust Estimation of the SUR Model," *The Canadian Journal of Statistics*, 28, 277-288.

Butler, R.W., Davies, P.L., and Jhun, M. (1993), "Asymptotics for the Minimum Covariance Determinant Estimator," *The Annals of Statistics,* 21, 1385–1400.

Charnes, A., Cooper, W.W., and Rhodes, E., (1981), "Evaluating Program and Managerial Efficiency: An Application of Data Envelopment Analysis to Program Follow Through," *Management Science,* 27, 668–697.

Coakley, C.W. and Hettmansperger, T.P. (1993), "A Bounded Influence, High Breakdown, Efficient Regression Estimator," *Journal of the American Statistical Association,* 88, 872-880.

Croux, C., and Haesbroeck, G. (1999), "Influence Function and Efficiency of the Mininmum Covariance Determinant Scatter Matrix Estimator," *Journal of Multivariate Analysis*, 71, 161–190.

Croux, C., Rousseeuw, P.J., and Hössjer, O. (1994), "Generalized S-Estimators," *Journal of the American Statistical Association,* 89, 1271-1281.

Donoho, D.L., and Huber, P.J. (1983), "The Notion of Breakdown Point," in *A Festschrift for Erich Lehmann* (P.J. Bickel, K.A. Doksum and J.L. Hodges, eds.), Belmont, Wadsworth, pp 157-184.

Grübel, R. (1988), "A Minimal Characterization of the Covariance Matrix," *Metrika*, 35, 49–52.

Hampel, F.R., Ronchetti E.M., Rousseeuw P.J., and Stahel W.A. (1986), *Robust Statistics: The Approach Based on Influence Functions,* John Wiley and Sons, New York.

Hössjer, O. (1994), "Rank-Based Estimates in the Linear Model With High Breakdown Point," *Journal of the American Statistical Association,* 89, 149-158.

Johnson, N.L., and Kotz, S. (1972), *Distributions in Statistics: Continuous Multivariate Distributions,* John Wiley and Sons, New York.

Koenker, R., and Portnoy, S. (1990), "M Estimation of Multivariate Regressions," *Journal of the American Statistical Association*, 85, 1060-1068.

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.

Marrona, R.A., and Yohai, V.J. (1997), "Robust Estimation in Simultaneous Equations Models," *Journal of Statistical Planning and Inference*, 57, 233-244.

Ollila, E., Hettmansperger, T.P., and Oja, H. (2002), "Estimates of Regression Coefficients Based on Sign Covariance Matrix," to appear in *Journal of the Royal Statistical Society, Series B.*

Ollila, E., Oja, H., and Koivunen, V. (2001), "Estimates of Regression Coefficients Based on Rank Covariance Matrix," submitted.

Rousseeuw, P.J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, 871–880.

Rousseeuw, P.J., and Leroy, A.M. (1987), *Robust Regression and Outlier Detection,* Wiley-Interscience, New York.

Rousseeuw, P.J., Van Aelst, S., Van Driessen, K., and Agulló, J. (2000), "Robust Multivariate Regression," submitted.

Rousseeuw, P.J., and Van Driessen K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics,* 41, 212–223.

Rousseeuw, P.J., and van Zomeren, B.C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association*, 85, 633–651.

Simpson, D.G., Ruppert, D., and Carroll, R.J. (1992), "On One-Step GM Estimates and Stability of Inferences in Linear Regression," *Journal of the American Statistical Association,* 87, 439-450.