

The Multivariate Least Trimmed Squares Estimator

Jose Agulló

*Departamento de Fundamentos del Análisis Económico, University of Alicante,
E-03080, Alicante, Spain*

Christophe Croux

*Faculty of Economics and Applied Economics, ORSTAT Research Center,
KULeuven, Naamsestraat 69, B-3000 Leuven, Belgium.*

Stefan Van Aelst *

*Dept. of Applied Mathematics and Computer Science, Ghent University
(UGENT), Krijgslaan 281 S9, B-9000 Gent, Belgium.*

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 estimator.

Key words: Multivariate Regression, Breakdown Point, Influence Function,
Minimum Covariance Determinant Estimator.

1 Introduction

Consider the multivariate regression model

$$y_i = \mathcal{B}^t x_i + \varepsilon_i \quad (1)$$

$i = 1, \dots, n$ with $x_i = (x_{i1}, \dots, x_{ip})^t \in \mathbb{R}^p$ and $y_i = (y_{i1}, \dots, y_{iq})^t \in \mathbb{R}^q$. The matrix $\mathcal{B} \in \mathbb{R}^{p \times q}$ contains the regression coefficients. The error terms $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. with zero center and as covariance 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 ($x_i = 1$). If the last regressor equals one, that is $x_{ip} = 1$ for all $1 \leq i \leq n$, we obtain a regression model with intercept. Denote the entire sample $Z_n = \{(x_i, y_i); i = 1, \dots, n\}$ and write $X = (x_1, \dots, x_n)^t$ for the design matrix and $Y = (y_1, \dots, y_n)^t$ for the matrix of responses. The classical estimator for \mathcal{B} is the least-squares (LS) estimator \mathcal{B}_{LS} which is given by

$$\hat{\mathcal{B}}_{LS} = (X^t X)^{-1} X^t Y \quad (2)$$

* Corresponding author.

Email addresses: `agullo@merlin.fae.ua.es` (Jose Agulló),
`Christophe.Croux@econ.kuleuven.ac.be` (Christophe Croux),
`Stefan.VanAelst@UGent.be` (Stefan Van Aelst).

while Σ is unbiasedly estimated by

$$\hat{\Sigma}_{LS} = \frac{1}{n-p}(Y - X\hat{\mathcal{B}}_{LS})^t(Y - X\hat{\mathcal{B}}_{LS}). \quad (3)$$

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 in [17] in the context of simultaneous equations models. Note that a simultaneous regression model is more general than model (1), since it allows for different regressors in each equation. Koenker and Portnoy [13] apply a regression M-estimator to each coordinate of the responses and Bai et al. [1] 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 in [18,19] using sign and rank based covariance matrices and in [22] using the Minimum Covariance Determinant estimator [20]. These methods mainly focus on random designs. Our approach will be more general, since it will be based only on the covariance matrix of the residuals, instead of on the covariance matrix of the joint distribution. Therefore, our method is well suited both for fixed and random designs.

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 reweighted MLTS estimator. Section 8 presents simulation results. Simulations have been done to compare the performance and robustness of the MLTS estimator with other alternatives. Section 9 concludes and 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. By taking the determinant, the correlation between the different components of the error term is taken into account. Note that the resulting estimator will simply be the LS-estimator computed from the optimal subset. The definition of the estimator is reminiscent of that of the MCD location/scatter estimator [20], and reduces to it in case of a multivariate regression model with only an intercept, where $X = (1, \dots, 1)^t \in \mathbb{R}^n$. 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. This has recently been investigated for S-estimators [2,25] and τ -estimators [7].

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 $Z_n = \{(x_i, y_i); i = 1, \dots, n\} \subset \mathbb{R}^{p+q}$ and for any $\mathcal{B} \in \mathbb{R}^{p \times q}$ denote $r_i(\mathcal{B}) = y_i - \mathcal{B}^t x_i$ the corresponding residuals. Let $\mathcal{H} = \{H \subset \{1, \dots, n\} | \#H = h\}$ 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 $\mathcal{B} \in \mathbb{R}^{p \times q}$ denote

$$\text{cov}(H, \mathcal{B}) := \frac{1}{h} \sum_{j \in H} (r_j(\mathcal{B}) - \bar{r}_H(\mathcal{B}))(r_j(\mathcal{B}) - \bar{r}_H(\mathcal{B}))^t, \quad (4)$$

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*

$$\hat{\mathcal{B}}_{MLTS}(Z_n) = \hat{\mathcal{B}}_{LS}(\hat{H}) \text{ where } \hat{H} \in \underset{H \in \mathcal{H}}{\text{argmin}} \det \hat{\Sigma}_{LS}(H) \quad (5)$$

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

$$\hat{\Sigma}_{MLTS}(Z_n) = c_\alpha \hat{\Sigma}_{LS}(\hat{H}), \quad (6)$$

where c_α is a consistency factor.

Note that if the minimization problem has more than one solution, in which case we look at $\underset{H}{\text{argmin}} \det \hat{\Sigma}_{LS}(H)$ 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 regression estimator. The

accompanying estimator of Σ is biased, however, due to the division by n in (4) instead of $n - p$ for the unbiased estimator in (3). Throughout the text we will suppose that no h points of the dataset $Z_n = \{(x_i, y_i); i = 1, \dots, n\} \subset \mathbb{R}^{p+q}$ are lying in the same subspace of \mathbb{R}^{p+q} . Formally, this means that for all $\beta \in \mathbb{R}^p$, $\gamma \in \mathbb{R}^q$, it holds that

$$\#\{(x_i, y_i) \mid \beta^t x_i + \gamma^t y_i = 0\} < h \quad (7)$$

unless if β and γ are both zero vectors.

For datasets satisfying condition (7) we now give two equivalent characterizations of the MLTS estimator. First we 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. For any $\mathcal{B} \in \mathbb{R}^{p \times q}$, let us 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 $\hat{H} \in \underset{H \in \mathcal{H}}{\text{argmin}} \det \text{Cov}_0(H, \mathcal{B})$. The residual covariance matrices we consider are thus centered at zero. (If we work with a model with intercept it can be shown that ‘‘Cov₀’’ may be replaced by the usual sample covariance matrix of the residuals.)

Proposition 1 *With the notations above, for datasets satisfying (7), we have that*

$$\underset{\mathcal{B}}{\text{argmin}} \det \text{MCD}_q(\mathcal{B}) = \{\hat{\mathcal{B}}_{LS}(\hat{H}) \mid \hat{H} \in \underset{H \in \mathcal{H}}{\text{argmin}} \det \hat{\Sigma}_{LS}(H)\} \quad (8)$$

Proposition 1 shows that any \mathcal{B} which minimizes the determinant of the MCD scatter estimate of its residuals is also a solution of (5). In the case of unique solutions, which we have almost surely if we sample from a continuous distri-

bution, we can rewrite (8) as

$$\hat{\mathcal{B}}_{MLTS}(Z_n) = \underset{\mathcal{B}}{\operatorname{argmin}} \det \operatorname{MCD}_q(\mathcal{B}). \quad (9)$$

For the residual scatter estimator we have

$$\hat{\Sigma}_{MLTS}(Z_n) = c_\alpha \operatorname{MCD}_q(\hat{\mathcal{B}}_{MLTS}(Z_n)) \quad (10)$$

A third characterization of the MLTS is based on the distances of the residuals. For any $\mathcal{B} \in \mathbb{R}^{p \times q}$ and $\Sigma \in \operatorname{PDS}(q)$, the class 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 $d_{1:n}(\mathcal{B}, \Sigma) \leq \dots \leq d_{n:n}(\mathcal{B}, \Sigma)$ the ordered sequence of distances of the residuals. Then the MLTS estimator can also be obtained in the following way.

Proposition 2 *Consider*

$$\underset{\mathcal{B}, \Sigma; |\Sigma|=1}{\operatorname{argmin}} \sum_{j=1}^h d_{j:n}^2(\mathcal{B}, \Sigma) \quad (11)$$

where the minimum is over all $\mathcal{B} \in \mathbb{R}^{p \times q}$ and $\Sigma \in \operatorname{PDS}(q)$ with $\det \Sigma = 1$ (denoted as $|\Sigma| = 1$). Then for datasets satisfying (7) it holds that

$$\left\{ \tilde{\mathcal{B}} \mid (\tilde{\mathcal{B}}, \tilde{\Sigma}) \in \underset{\mathcal{B}, \Sigma; |\Sigma|=1}{\operatorname{argmin}} \sum_{j=1}^h d_{j:n}^2(\mathcal{B}, \Sigma) \right\} = \{ \hat{\mathcal{B}}_{LS}(\hat{H}) \mid \hat{H} \in \underset{H}{\operatorname{argmin}} \det \hat{\Sigma}_{LS}(H) \}. \quad (12)$$

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 (5). In the

case of unique solutions, Proposition 2 yields

$$\hat{\mathcal{B}}_{MLTS}(Z_n) = \operatorname{argmin}_{\mathcal{B}, \Sigma; |\Sigma|=1} \sum_{j=1}^h d_{j:n}^2(\mathcal{B}, \Sigma), \quad (13)$$

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 (11) reduces to $\operatorname{argmin}_{\mathcal{B}} \sum_{j=1}^h r_{j:n}^2(\mathcal{B})$, with $r_{1:n}(\mathcal{B}) \leq \dots \leq r_{n:n}(\mathcal{B})$ 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 (LTS) estimator [20]. 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 [6]. Formally, it is defined as

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

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 $Z_n \subset \mathbb{R}^{p+q}$ denote $k(Z_n)$ the maximal number of observations of Z_n lying on a same subspace of \mathbb{R}^{p+q} . Since we required that Z_n satisfies (7), we have $k(Z_n) < h$. We now have the following theorem.

Theorem 1 For any dataset $Z_n \subset \mathbb{R}^{p+q}$ satisfying (7) with $q > 1$ it holds that

$$\varepsilon_n^*(\hat{\mathcal{B}}_{MLTS}, Z_n) = \frac{\min(n - h + 1, h - k(Z_n))}{n}. \quad (14)$$

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 subspace of \mathbb{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 1: For a regression model with intercept we can explicitly write $x_i = (u_i^t, 1)^t$ with $u_i = (x_{i1}, \dots, x_{i,p-1})^t$ in (1). In this case, the last row of \mathcal{B} is the intercept vector $(\mathcal{B}^0)^t$ and the matrix formed by the $p - 1$ first rows of \mathcal{B} is the slope matrix \mathcal{B}^1 . The previous result holds for both the slope matrix and intercept vector.

Corollary 1 For datasets $Z_n \subset \mathbb{R}^{p+q}$ with $q > 1$ and satisfying (7) with $\beta \neq 0$ it holds that

$$\varepsilon_n^*(\hat{\mathcal{B}}_{MLTS}^1, Z_n) = \varepsilon_n^*(\hat{\mathcal{B}}_{MLTS}^0, Z_n) = \frac{\min(n - h + 1, h - k(Z_n))}{n}. \quad (15)$$

Remark 2: 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 2 Denote $k'(Z_n)$ the maximal number of $x_j \in \{x_i; i = 1, \dots, n\}$

lying on a subspace of \mathbb{R}^p . Then for any dataset $Z_n \subset \mathbb{R}^{p+1}$ with $k'(Z_n) < h$ it holds that

$$\varepsilon_n^*(\hat{\mathcal{B}}_{LTS}, Z_n) = \frac{\min(n - h + 1, h - k'(Z_n))}{n}. \quad (16)$$

If Z_n comes from a continuous distribution H 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 in [11]. In this case any $[(n + p)/2] \leq h \leq [(n + p + 1)/2]$ gives the maximal breakdown point.

Remark 3: In the case $p = 1$ and $x_i = 1$ Theorem 1 gives the following result for the breakdown point of the MCD estimator of multivariate location and scatter.

Corollary 3 Consider $Y_n = \{y_1, \dots, y_n\} \subset \mathbb{R}^q$ and denote $k''(Y_n)$ the maximal number of $y_j \in Y_n$ lying on a hyperplane of \mathbb{R}^q . Then for any dataset $Y_n \subset \mathbb{R}^q$ with $k''(Y_n) < h$ it holds that

$$\varepsilon_n^*(\hat{\mu}_{MCD}, Y_n) = \frac{\min(n - h + 1, h - k''(Y_n))}{n}. \quad (17)$$

If Y_n comes from a continuous distribution F then almost surely $k''(Y_n) = q - 1$ which yields $\varepsilon_n^*(\hat{\mu}_{MCD}, Y_n) = \varepsilon_n^*(\hat{\Sigma}_{MCD}, Y_n) = \min(n - h + 1, h - q + 1)/n$. Therefore, any $[(n + q)/2] \leq h \leq [(n + q + 1)/2]$ gives the maximal breakdown point.

4 Algorithm

Recently, a fast algorithm has been developed to compute the MCD location and scatter estimator [23]. 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 $d_{1:n}(\hat{\mathcal{B}}_1, \hat{\Sigma}_1) \leq \dots \leq d_{h:n}(\hat{\mathcal{B}}_1, \hat{\Sigma}_1)$. 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, as in [23], 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 $\{1, \dots, n\}$ and compute the corresponding least squares estimates $\hat{\mathcal{B}}_m := \hat{\mathcal{B}}_{LS}(J_m)$ 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 $i = 1, \dots, n$ 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 determinant 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 in [23]. The total number of random starts m should be large enough such that the probability of finding the global minimum is high. In our experience using $m = 1000$ random starts is often sufficient. However, in higher dimensions more random starts may be required to get a stable solution. See e.g. [10] for more discussion on the number of starting points in resampling algorithms.

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

$$\mathcal{D}_K(\alpha) = \{A \mid A \subset \mathbb{R}^{p+q} \text{ measurable and bounded with } P_K(A) = 1 - \alpha\}. \quad (18)$$

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

$$P_K(\beta^t x = 0) < 1 - \alpha \text{ for all } \beta \in \mathbb{R}^p \setminus \{0\}. \quad (19)$$

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

$$\mathcal{B}_A(K) = \left(\int_A x x^t dK(x, y) \right)^{-1} \int_A x y^t dK(x, y) \quad (20)$$

and

$$\Sigma_A(K) = \frac{\int_A (y - \mathcal{B}_A(K)^t x)(y - \mathcal{B}_A(K)^t x)^t dK(x, y)}{1 - \alpha}. \quad (21)$$

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

$$\mathcal{B}_{MLTS}(K) = \mathcal{B}_{\hat{A}}(K) \text{ and } \Sigma_{MLTS}(K) = c_\alpha \Sigma_{\hat{A}}(K). \quad (22)$$

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 [4] to ensure that the set $\mathcal{D}_K(\alpha)$ is non-empty.

Now consider the multivariate regression model

$$y = \mathcal{B}^t x + \varepsilon \quad (23)$$

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 assume 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 $\Sigma \in PDS(q)$. Furthermore, 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. Note that we do not need that F_Σ has finite second moments, but if second moments exist, then Σ is proportional to the covariance matrix of the distribution. The distribution of $z = (x, y)$ is denoted by H . A regularity condition (to avoid degenerate situations) on the model distribution H is that

$$P_H(\beta^t x + \gamma^t y = 0) < 1 - \alpha \quad (24)$$

for all $\beta \in \mathbb{R}^p$ and $\gamma \in \mathbb{R}^q$ not both equal to zero at the same time. If $\alpha = 0$ this regularity condition means that the distribution H is not com-

pletely 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 $1 - \alpha$. Note that condition (24) implies condition (19) because γ can be put equal to zero. We first give the following proposition which says that the MLTS solution can always be taken as a cylinder.

Lemma 1 *Consider a distribution H satisfying (24) and an MLTS solution $\hat{A} \in D_H(\alpha)$. For any $(x, y) \in \mathbb{R}^{p+q}$ 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 $\mathcal{E} = \{(x, y) \in \mathbb{R}^{p+q}; d^2(x, y) \leq D_\alpha^2\}$ 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 (22) for some well chosen constant c_α are Fisher-consistent for the parameters \mathcal{B} and Σ .

Theorem 3 *Denote*

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

where $F_0 = F_{I_q}$ is the central error distribution and $q_\alpha = F_*^{-1}(1 - \alpha)$ with $F_*(t) = P_{F_0}(U^t U \leq t)$. 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 to obtain 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 the constant c_α reduces to $c_\alpha = (1 - \alpha)/F_{\chi_{q+2}^2}(q_\alpha)$

with $q_\alpha = \chi_{q,1-\alpha}^2$, the upper α percent point of the χ^2 distribution with q degrees of freedom and $F_{\chi_{q+2}^2}$ the cumulative distribution function of the χ^2 distribution with $q + 2$ degrees of freedom.

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 [9].) 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 and if $E(\|x\|^2) < \infty$, we have that*

$$IF(z; \mathcal{B}_{MLTS}, H_0) = E_{H_0}[xx^t]^{-1} \frac{xy^t}{-2c_2} I(\|y\|^2 \leq q_\alpha) \quad (25)$$

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 (25) 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 [5, 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 [3,4]. 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

$$IF(z; \mathcal{B}_{MLTS}, H_0) = E_{H_0}[xx^t]^{-1} x IF(y; T_q, F_0)^t. \quad (26)$$

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

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

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 [4].

Remark 4: For models with intercept we can explicitly write $x = (u^t, 1)^t$ with $u = (x_1, \dots, x_{p-1})^t$ in (23). In this case, the last row of \mathcal{B} is the intercept vector $(\mathcal{B}^0)^t$ and the matrix formed by the $p - 1$ first rows of \mathcal{B} is the slope matrix \mathcal{B}^1 . Denote $\mu_u = E[u]$ and $\Sigma_u = E[(u - \mu_u)(u - \mu_u)^t]$, it then follows immediately from (25) and (26) that

$$\begin{aligned} IF(z; \mathcal{B}_{MLTS}^1, H_0) &= \Sigma_u^{-1}(u - \mu_u) \frac{y^t}{-2c_2} I(\|y\|^2 \leq q_\alpha) \\ &= \Sigma_u^{-1}(u - \mu_u) IF(y; T_q, F_0)^t \\ IF(z; \mathcal{B}_{MLTS}^0, H_0) &= (y - y(u - \mu_u)^t \Sigma_u^{-1} \mu_u) \frac{I(\|y\|^2 \leq q_\alpha)}{-2c_2} \\ &= IF(y; T_q, F_0) - IF(z; \mathcal{B}_{MLTS}^1, H_0)^t \mu_u. \end{aligned}$$

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. [9]). 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 $(d_1 d_3 \times d_2 d_4)$ matrix with $d_1 d_2$ blocks of size $(d_3 \times d_4)$. For $1 \leq j \leq d_1$ and $1 \leq k \leq d_2$ 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 (26) implies that

$$ASV(\mathcal{B}_{MLTS}, H_0) = D_{p,q}(\text{diag}(ASV(T_q, F)) \otimes \Sigma_x^{-1}) \quad (27)$$

where the commutation matrix $D_{p,q}$ is a $(pq \times pq)$ matrix consisting of pq blocks of size $(q \times p)$. For $1 \leq l \leq p$ and $1 \leq m \leq q$ 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 (27) it follows that for every $1 \leq i \leq p$ and $1 \leq j \leq q$ the asymptotic covariance matrix of $(\mathcal{B}_{MLTS})_{ij}$ is given by $\Delta_{ji} \Sigma_x^{-1} ASV((T_q)_j, F)$ which implies

that the asymptotic variance of $(\mathcal{B}_{MLTS})_{ij}$ equals

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

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

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

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 $ASV((\mathcal{B}_{LS})_{ij}, H_0) = ASV(M_j, F)$ 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.

	MLTS					RMLTS				
α	$q = 2$	$q = 3$	$q = 5$	$q = 10$	$q = 30$	$q = 2$	$q = 3$	$q = 5$	$q = 10$	$q = 30$
0.25	0.403	0.466	0.531	0.597	0.664	0.941	0.954	0.965	0.974	0.982
0.5	0.153	0.204	0.262	0.327	0.398	0.934	0.951	0.963	0.973	0.982

Table 1

Asymptotic relative efficiency of the MLTS and RMLTS estimators w.r.t. the least squares estimator at the normal distribution for several values of q .

7 Reweighting

The efficiency of MLTS can be quite low, as can be seen from Table 1. Therefore, we now introduce a one-step reweighted estimator that improves the performance of the 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

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

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 [21] we used $\delta = 0.01$ and $q_\delta = \chi_{q, 1-\delta}^2$ the corresponding quantile of the χ^2 distribution with q degrees of freedom. In the case of multivariate normal errors we have $c_\delta = (1 - \delta) / F_{\chi_{q+2}^2}(q_\delta)$.

For any distribution K satisfying (19) the RMLTS functionals can be written as

$$\mathcal{B}_{RMLT}(K) = \left(\int_J xx^t dK(x, y) \right)^{-1} \int_J xy^t dK(x, y) \quad (28)$$

$$\Sigma_{RMLTS}(K) = c_\delta \frac{\int_J (y - \mathcal{B}_{RMLTS}(K))^t x (y - \mathcal{B}_{RMLTS}(K))^t x^t dK(x, y)}{1 - \delta} \quad (29)$$

where $J = \{(x, y) : d^2(\mathcal{B}_{MLTS}(K), \Sigma_{MLTS}(K)) \leq q_\delta\}$. Following [22] we obtain for any model distribution H_0 as in Theorem 4 that

$$IF(z, \hat{\mathcal{B}}_{RMLTS}, H_0) = \left(1 + \frac{2d_2}{1 - \delta}\right) IF(z, \hat{\mathcal{B}}_{MLTS}, H_0) + \frac{E_{H_0}[xx^t]^{-1}}{1 - \delta} xy^t I(\|y\|^2 \leq q_\delta) \quad (30)$$

where the constant d_2 is the same as c_2 in Theorem 4 but with α replaced by δ . Similarly, as for the initial MLTS, this influence function is bounded in y but unbounded in x . Good leverage points can have a high effect on the MLTS estimator but bad leverage points will give a zero influence.

Remark 1: The influence function of the reweighted MCD location estimator T_q^1 at a q -dimensional spherical distribution F_0 equals

$$IF(y, T_q^1, F_0) = \left(1 + \frac{2d_2}{1 - \delta}\right) IF(y, T_q, F_0) + \frac{y}{1 - \delta} I(\|y\|^2 \leq q_\delta).$$

Therefore, the influence function of \mathcal{B}_{RMLTS} can be rewritten as

$$IF(z; \mathcal{B}_{RMLTS}, H_0) = E_{H_0}[xx^t]^{-1} x IF(y; T_q^1, F_0)^t. \quad (31)$$

Remark 2: It can also be shown that

$$IF(z; \Sigma_{RMLTS}, H_0) = IF(y, C_q^1, F_0)$$

where C_q^1 is the q -dimensional reweighted MCD scatter estimator (RMCD). The influence function of the RMCD scatter estimator at elliptical distributions can be obtained from [4].

Analogous to (27) we obtain from (31) that the asymptotic variance-covariance

matrix of \mathcal{B}_{RMLTS} equals

$$ASV(\mathcal{B}_{RMLTS}, H_0) = D_{p,q}(\text{diag}(ASV(T_q^1, F)) \otimes \Sigma_x^{-1}). \quad (32)$$

Hence, the asymptotic variances and covariances of $(\mathcal{B}_{MLTS})_{ij}$ are

$$\begin{aligned} ASV((\mathcal{B}_{RMLTS})_{ij}, H_0) &= (\Sigma_x^{-1})_{ii} ASV((T_q^1)_j, F) \\ ASC((\mathcal{B}_{RMLTS})_{ij}, (\mathcal{B}_{RMLTS})_{i'j}, H_0) &= (\Sigma_x^{-1})_{ii'} ASV((T_q^1)_j, F) \quad \text{for } i \neq i' \end{aligned}$$

and all other asymptotic covariances (for $j' \neq j$) equal 0. The asymptotic relative efficiency of the RMLTS estimator at the model distribution H_0 with respect to the least squares estimator becomes

$$ARE((\mathcal{B}_{RMLTS})_{ij}, H_0) = ARE((T_q^1)_j, F_0)$$

for all $1 \leq i \leq p$ and $1 \leq j \leq q$, the asymptotic relative efficiency of the q -dimensional RMCD location estimator at the error distribution F_0 . For the normal distribution these relative efficiencies are also given in Table 1. Note that reweighting the MLTS improves its efficiency a lot. Moreover, the difference in efficiency between RMLTS based on the initial MLTS with 25% BDP and 50% BDP is very small and vanishes with increasing value of q .

8 Finite-sample simulations

8.1 Finite-sample performance

In this section we investigate the finite-sample performance of the MLTS and RMLTS estimators and compare it with other robust multivariate regression estimators. To this end, we performed the following simulations. We generated $m = 1000$ regression datasets of size $n = 100$, $n = 300$ and $n = 500$. We will

discuss results for $p = q = 3$ and $p = 10, q = 5$ in this paper. We set the p th 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. [12, p. 134].)

W.l.o.g. we took $\mathcal{B} = 0$ in the multivariate regression model. The response variables were generated from the multivariate standard normal distribution $N(0, I_q)$ or multivariate Cauchy distribution.

To compare the performance of MLTS and RMLTS with other 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

$$\text{MSE}(T) = n \text{average}_l (T^{(l)} - \theta)^2 \quad l = 1, \dots, m$$

where θ is the true value of the parameter and $T^{(l)}$ are the estimates for the simulated dataset $Z^{(l)}, l = 1, \dots, m$. The MSE of a slope matrix estimator $\hat{\mathcal{B}}^1$ is then defined as

$$\text{MSE}(\hat{\mathcal{B}}^1) = \text{average}_{1 \leq j \leq p-1, 1 \leq k \leq q} (\text{MSE}(\hat{\mathcal{B}}_{jk}^1))$$

and similarly for the intercept vector. Throughout the paper the results for the slope will be shown and the results for the intercept will be omitted because they yield the same conclusions.

Table 2 shows the MSE of the MLTS and RMLTS estimators, the biweight S-estimator [15,25], and the MCD and LR-weighted MCD (LRMCD) regression estimators [22] all with 50% breakdown point. Results for the multivariate M-estimator proposed in [13] are included as well. The M-estimator uses the Huber ψ function with tuning constant that yields 95% efficiency at the model with normal errors. The MLTS estimator was computed with the algorithm outlined in section 4. The MCD regression algorithm uses the FAST-MCD algorithm [23]. The S-estimator was computed using local improvement steps from the MLTS which generalizes the S-estimator algorithm of multivariate location and scatter [26] to multivariate regression. This algorithm is faster than the resampling approach proposed in [24], thus the MLTS is a useful initial estimator for computing S-estimators. This choice of algorithms implies that all high-breakdown estimators have the same time complexity. Note however, that MCD-regression requires computation of the MCD in $p + q$ dimensions while MLTS mainly requires computations in q dimensions. Hence, for fixed dimensions p and q , the MLTS will be faster to compute than MCD regression. From Table 2 we see that the reweighting step largely improves the performance of the initial MLTS estimator. The coordinatewise M-estimator performs best followed closely by the RMLTS, S, and LRMCD estimators. Moreover, we see that except for MCD regression, results obtained for the asymmetric exponential carriers are comparable to those obtained for normal carriers. This confirms that contrary to MCD regression the efficiency of MLTS does not depend on the distribution of the carriers when the carriers are un-

correlated. Under $n = \infty$ the asymptotic variance of the estimators for normal distributions is listed. We see that the mean squared error at normal samples converges to the corresponding asymptotic variance but convergence for MCD regression in low dimensions is very slow. Moreover, the MSE of MLTS for sample size $n = 100$ is already comparable to the asymptotic variance which indicates that the MLTS algorithm provides good solutions.

In Figure 1a we investigate the performance of the estimators at long tailed carrier (CAU) distributions. The results for $p = q = 3$ are shown in the left panel while the right panel shows the results for $p = 10, q = 5$. From these plots we see that coordinatewise M and S-estimators show the best performance followed closely by RMLTS. MCD and LRMCD regression perform worse than the initial MLTS estimator in this setting. Figure 1b compares the performance of the estimators at long tailed error (CAU) distributions. Now the coordinatewise M-estimator is clearly worse than all others. The S-estimator performs best while MLTS, RMLTS and LRMCD show similar behavior. Overall, we can conclude that the biweight S-estimator and RMLTS estimators show stable good performance in all cases considered.

8.2 *Finite-sample robustness*

To study the finite-sample robustness of the MLTS estimator we carried out simulations with contaminated datasets. To generate contaminated datasets 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 $N(\lambda\sqrt{\chi_{p-1, .99}^2}, 1.5)$ and the q dependent variables were generated from $N(\kappa\sqrt{\chi_{q, .99}^2}, 1.5)$. Both λ and κ took values in

		$n = 100$		$n = 300$		$n = 500$		$n = \infty$
dimensions	method	NOR	EXP	NOR	EXP	NOR	EXP	NOR
p=3 q=3	M	1.10	1.19	1.09	1.08	1.07	1.07	1.05
	MLTS	4.09	4.99	4.42	4.77	4.85	4.83	4.90
	RMLTS	1.79	2.25	1.21	1.30	1.17	1.18	1.05
	S	1.52	1.70	1.42	1.49	1.45	1.43	1.39
	MCD	7.94	21.25	8.48	20.74	9.42	21.22	7.52
	LRMCD	1.56	2.44	1.12	1.22	1.12	1.14	1.05
p=10 q=5	M	1.21	1.28	1.10	1.11	1.08	1.08	1.05
	MLTS	3.17	3.99	3.52	3.94	3.63	3.59	3.82
	RMLTS	2.35	3.00	1.31	1.42	1.19	1.21	1.04
	S	1.33	1.52	1.25	1.29	1.24	1.21	1.18
	MCD	4.51	6.50	4.33	6.21	4.28	6.12	4.24
	LRMCD	3.33	4.42	1.77	1.67	1.44	1.35	1.04

Table 2

MSE of the slope at normal (NOR) or exponential (EXP) carrier distributions and normal error distribution.

$\{0, 0.5, 1, 1.5, 2, 3, 4, 5\}$. If $\lambda = 0$ and $\kappa > 0$, we obtain vertical outliers. On the other hand, if $\lambda > 0$ but $\kappa = 0$ we have good leverage points. Finally, if both $\lambda > 0$ and $\kappa > 0$, this yields bad leverage points. Note that large values of λ and κ produce extreme outliers while small values produce intermediate

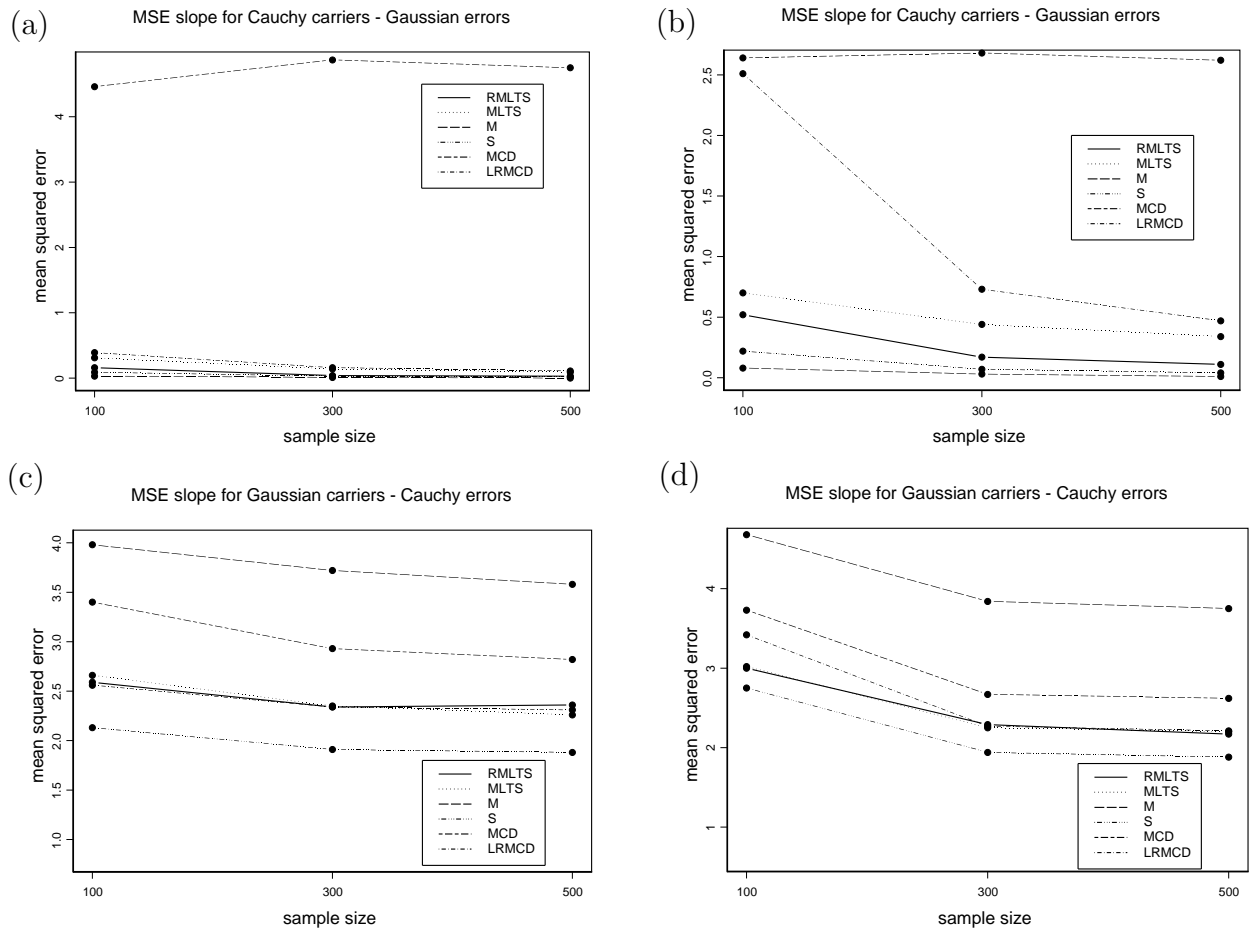


Fig. 1. MSE of the slope at Cauchy carrier distribution and Gaussian error distribution (top) and Gaussian carrier distribution and Cauchy error distribution (bottom). Left panels are for $p = q = 3$ and right panels are for $p = 10, q = 5$.

outliers.

In Figure 2 we show for each value of λ the maximal value of MSE obtained over all possible values of κ . We show results for datasets of size $n = 100$ from the model with $p = 10$ and $q = 5$ and Gaussian errors. Results for other sample sizes and dimensions were similar. The top panels in Figure 2 shows results for Gaussian carriers, while the bottom panels are for Cauchy carriers. The left plots show results for coordinatewise M and 50% breakdown estimators while the right plots are for 25% breakdown estimators.

From Figure 2 we immediately see that the coordinatewise M-estimator can produce extremely high MSE when the data contains contamination. The top panels show that MCD and LRMCD regression perform extremely well for data with a joint Gaussian distribution. This is no surprise because this approach is fine-tuned for joint elliptical distributions. However, the MSE of MLTS and RMLTS is also reasonably small. The bottom panels reveal that when the data is not jointly elliptical as is the case with Cauchy carriers, then the MSE of MCD and LRMCD regression can become very large. On the other hand, the MSE of MLTS and RMLTS are lower for Cauchy carriers than for Gaussian carriers. Finally, comparing left and right panels we see that the S-estimator has a very low MSE when the fraction of contamination is small compared to the BDP of the estimator. However, when the contamination fraction is closer to the BDP as in the right panels, the S-estimator can become affected more heavily, especially by intermediate outliers.

Overall, we see that MLTS and RMLTS always have a reasonably low MSE in the presence of outliers which confirms the robustness of these estimators. Furthermore, in most cases RMLTS improves the MSE of the initial MLTS and often this improvement is substantial. To summarize, RMLTS has shown good performance under uncontaminated data as well as stable robust behavior for contaminated data.

9 Conclusions

In this paper we have introduced the multivariate least trimmed squares estimator. We have given three equivalent definitions of the MLTS estimator which allow us to completely investigate and explain the behaviour of the

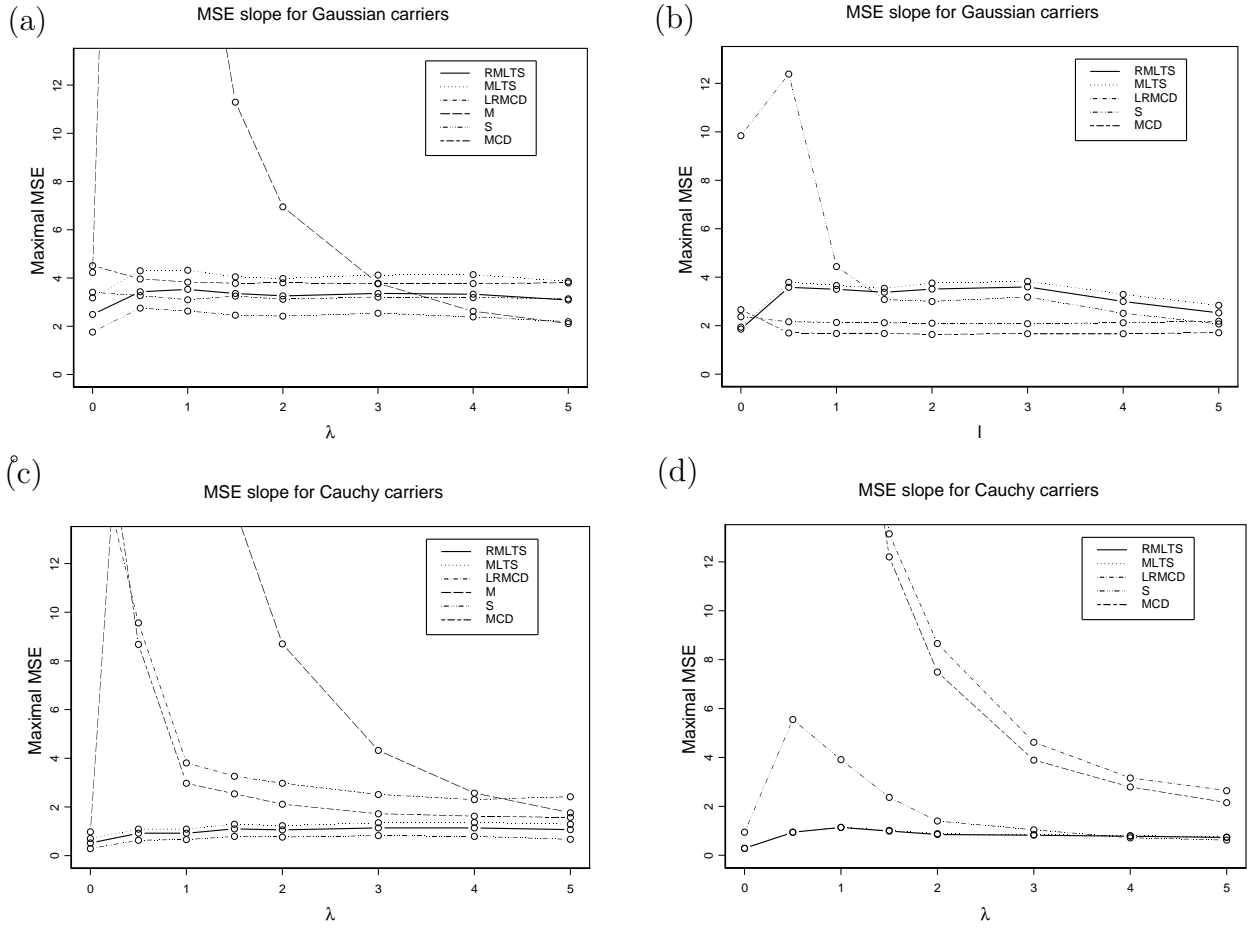


Fig. 2. MSE of the RMLTS, MLTS, M, MCD, LRMCD, and S estimators at contaminated data with normal and Cauchy carrier distributions. Left panels are results for BDP=50% and right panels for BDP=25%.

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 $h = \lceil (n + p + q + 1)/2 \rceil$ which yields the maximal breakdown point $\varepsilon_n^* \approx 50\%$ and $h \approx 0.75n$ which gives a 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 performance and robustness. We also investigated the one-step reweighted MLTS estimator. In all situations the RMLTS is similar or better than the initial MLTS estimator. Therefore, we recommend to use the one-step reweighted MLTS.

Another recent paper also introduced the Multivariate Least-trimmed Squares regression estimator [14]. In contrast to our work, this paper does not provide any theoretical results like consistency, influence functions, and asymptotic variances. The paper only contains an (incorrect) statement of the finite-sample breakdown point and proposes to compute the MLTS estimator by a feasible subset exchange algorithm, which is much less time-efficient than the procedure outlined in Section 4 of this paper. In our paper, we tried to give a complete analysis of the multivariate least squares estimator and its reweighted version.

Acknowledgments

We would like to thank the referees for their constructive comments. The research of Christophe Croux has been supported by the Research Fund K.U. Leuven and the Fund for Scientific Research-Flanders (Contract numbers G.0385.03 and G.0595.05). The research of Stefan Van Aelst has been supported by the Fund for Scientific Research - Flanders.

A Appendix

First, we show the following lemma which is a generalization of the characterization in [8] of the mean and covariance matrix of a multivariate distribution.

Lemma 2 *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 $b \in \mathbb{R}^{p+q}$ and Δ a positive definite and symmetric matrix of size q such that*

$$E_K[(y - b^t x)^t \Delta^{-1} (y - b^t x)] = q, \quad (\text{A.1})$$

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 \mathbb{R}^{p+q} , then Lemma 2 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.

Proof of Lemma 2. For ease of notation, let $\Sigma_{LS}(K) := \Sigma_{LS}$ and drop the subscript K . Put $u = \Sigma_{LS}^{-1/2} \varepsilon$. Then $E[(y - \mathcal{B}_{LS}^t x)^t \Sigma_{LS}^{-1} (y - \mathcal{B}_{LS}^t x)] = E[u^t u] = \text{tr } E[uu^t] = \text{tr}(\Sigma_{LS}^{-1/2} E[\varepsilon\varepsilon^t] \Sigma_{LS}^{-1/2}) = \text{tr } I_q = q$, so $(\mathcal{B}_{LS}, \Sigma_{LS})$ satisfies condition (A.1). Take any $b \in \mathbb{R}^{p+q}$ and any Δ a positive definite symmetric matrix of size q such that (A.1) holds. There exists an orthogonal matrix P and $\lambda_1 \geq \dots \geq \lambda_q > 0$ such that $\Delta = \Sigma_{LS}^{1/2} P \Lambda P^t \Sigma_{LS}^{1/2}$ where $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_q)$. Put $v = P^t \Sigma_{LS}^{-1/2} (y - b^t x)$. Then we obtain

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

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

$$\begin{aligned}
E[vv^t] &= P^t \Sigma_{LS}^{-1/2} E[(\varepsilon + (\mathcal{B}_{LS} - b)^t x)(\varepsilon + (\mathcal{B}_{LS} - b)^t x)^t] \Sigma_{LS}^{-1/2} P \\
&= P^t (I_q + \Sigma_{LS}^{-1/2} (\mathcal{B}_{LS} - b)^t E[xx^t] (\mathcal{B}_{LS} - b) \Sigma_{LS}^{-1/2}) P \\
&= I_q + ((\mathcal{B}_{LS} - b) \Sigma_{LS}^{-1/2} P)^t E[xx^t] ((\mathcal{B}_{LS} - b) \Sigma_{LS}^{-1/2} P). \tag{A.3}
\end{aligned}$$

Taking the diagonal elements of (A.3) and inserting them in (A.2) yields

$$q = \sum_{i=1}^q \lambda_i^{-1} + \sum_{i=1}^q \lambda_i^{-1} ((\mathcal{B}_{LS} - b) \Sigma_{LS}^{-1/2} P)_i^t E[xx^t] ((\mathcal{B}_{LS} - b) \Sigma_{LS}^{-1/2} P)_i \geq \sum_{i=1}^q \lambda_i^{-1}, \tag{A.4}$$

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

$$\sum_{i=1}^q \frac{1}{\lambda_i} \geq q \left(\prod_{i=1}^q \frac{1}{\lambda_i} \right)^{1/q} = q (\det \Lambda)^{-1/q} = q \left(\frac{\det \Sigma_{LS}}{\det \Delta} \right)^{1/q}. \tag{A.5}$$

From the last two inequalities (A.4) and (A.5) we see that $\det \Sigma_{LS} \leq \det \Delta$, showing already that $(\mathcal{B}_{LS}, \Sigma_{LS})$ solves the minimization problem.

Moreover, equality in (A.4) 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 (A.5) 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. \square

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 $\det \operatorname{MCD}_q(\mathcal{B})$. Take $\mathcal{B} \in \mathbb{R}^{p \times q}$ arbitrarily, then by definition of the MCD there exists a $H \in \mathcal{H}$ such that $\operatorname{MCD}_q(\mathcal{B}) = \operatorname{Cov}_0(H, \mathcal{B})$.

Using properties of traces, it follows that

$$\frac{1}{h} \sum_{j \in H} r_j(\mathcal{B}) (\operatorname{Cov}_0(H, \mathcal{B}))^{-1} r_j(\mathcal{B})^t = q. \tag{A.6}$$

Since the data satisfies condition (7), Lemma 2 can be applied:

$$\begin{aligned}
\det \operatorname{MCD}_q(\mathcal{B}) &= \det \operatorname{Cov}_0(H, \mathcal{B}) \geq \det \hat{\Sigma}_{LS}(H) \geq \det \hat{\Sigma}_{LS}(\hat{H}) = \det \operatorname{Cov}_0(\hat{H}, \hat{\mathcal{B}}_{LS}(\hat{H})) \\
&\geq \det \operatorname{MCD}_q(\hat{\mathcal{B}}_{LS}(\hat{H})),
\end{aligned}$$

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

On the other hand, take now $\tilde{\mathcal{B}} \in \underset{\mathcal{B}}{\text{argmin}} \det \text{MCD}_q(\mathcal{B})$. By definition of MCD , there exists a $\tilde{H} \in \mathcal{H}$ such that $\text{MCD}_q(\tilde{\mathcal{B}}) = \text{Cov}_0(\tilde{H}, \tilde{\mathcal{B}})$ and in particular $\det \text{Cov}_0(\tilde{H}, \tilde{\mathcal{B}}) \leq \det \text{Cov}_0(\tilde{H}, \hat{\mathcal{B}}_{LS}(\tilde{H}))$. But since (A.6) also holds for the pair $(\tilde{H}, \tilde{\mathcal{B}})$, the uniqueness part of Lemma 2 gives $\tilde{\mathcal{B}} = \hat{\mathcal{B}}_{LS}(\tilde{H})$. It then follows that for any other $H \in \mathcal{H}$ we have

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

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

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

$$\begin{aligned} \frac{1}{h} \sum_{j \in H} d_j^2(\hat{\mathcal{B}}_{LS}(H), \hat{\Sigma}_{LS}(H)) &= \frac{1}{h} \text{tr} \sum_{j \in H} \hat{\Sigma}_{LS}(H)^{-1} r_j(\hat{\mathcal{B}}_{LS}(H)) r_j(\hat{\mathcal{B}}_{LS}(H))^t \\ &= \text{tr} \hat{\Sigma}_{LS}(H)^{-1} \hat{\Sigma}_{LS}(H) = q. \end{aligned} \quad (\text{A.7})$$

We also have that

$$\sum_{j \in H} d_j^2(\hat{\mathcal{B}}_{LS}(H), \hat{\Sigma}_{LS}(H)) = (\det \hat{\Sigma}_{LS}(H))^{-1/q} \sum_{j \in H} d_j^2(\hat{\mathcal{B}}_{LS}(H), \tilde{\Sigma}_{LS}(H)) \quad (\text{A.8})$$

Combining (A.8) with (A.7) yields

$$\sum_{j \in H} d_j^2(\hat{\mathcal{B}}_{LS}(H), \tilde{\Sigma}_{LS}(H)) = hq \det \hat{\Sigma}_{LS}(H)^{1/q}. \quad (\text{A.9})$$

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

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

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 (A.8) and (A.9), 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 2 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

$$\sum_{j=1}^h d_{j:n}^2(\hat{\mathcal{B}}_{LS}(\hat{H}), \tilde{\Sigma}_{LS}(\hat{H})) = \sum_{j \in \hat{H}} d_j^2(\hat{\mathcal{B}}_{LS}(\hat{H}), \tilde{\Sigma}_{LS}(\hat{H})). \quad (\text{A.10})$$

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

$$\sum_{j=1}^h d_{j:n}^2(\mathcal{B}, \Sigma) < \sum_{j=1}^h d_{j:n}^2(\hat{\mathcal{B}}_{LS}(\hat{H}), \tilde{\Sigma}_{LS}(\hat{H})) \quad (\text{A.11})$$

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

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

Using (A.9) this implies that $\frac{1}{h} \sum_{j \in H_1} d_j^2(\mathcal{B}, \det \hat{\Sigma}_{LS}(H_1)^{1/q} \Sigma) < q$. Hence, there exists a constant $0 < c < 1$ such that $\frac{1}{h} \sum_{j \in H_1} d_j^2(\mathcal{B}, c \det \hat{\Sigma}_{LS}(H_1)^{1/q} \Sigma) = q$. From Lemma 2 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

$$\sum_{j=1}^h d_{j:n}^2(\mathcal{B}, \Sigma) \geq \sum_{j \in H_1} d_j^2(\hat{\mathcal{B}}_{LS}(H_1), \tilde{\Sigma}_{LS}(H_1)). \quad (\text{A.12})$$

From (A.10) and (A.12) it follows that the inequality (A.11) implies that

$$\sum_{j \in H_1} d_j^2(\hat{\mathcal{B}}_{LS}(H_1), \tilde{\Sigma}_{LS}(H_1)) < \sum_{j \in \hat{H}} d_j^2(\hat{\mathcal{B}}_{LS}(\hat{H}), \tilde{\Sigma}_{LS}(\hat{H})). \quad (\text{A.13})$$

But, using (A.9), 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

$\hat{H} \in \operatorname{argmin}_H \det \hat{\Sigma}_{LS}(H)$. Therefore, we conclude that

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

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

We now prove that for any $(\tilde{\mathcal{B}}, \tilde{\Sigma}) \in \operatorname{argmin}_{\mathcal{B}, \Sigma; |\Sigma|=1} \sum_{j=1}^h d_{j:n}^2(\mathcal{B}, \Sigma)$ there exists a $\tilde{H} \in \mathcal{H}$ such that $\tilde{\mathcal{B}} = \hat{\mathcal{B}}_{LS}(\tilde{H})$ and $\tilde{H} \in \operatorname{argmin}_H \det \hat{\Sigma}_{LS}(H)$. 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

$$\sum_{j=1}^h d_{j:n}^2(\tilde{\mathcal{B}}, \tilde{\Sigma}) = \sum_{j \in \tilde{H}} d_j^2(\tilde{\mathcal{B}}, \tilde{\Sigma}) \leq \sum_{j \in \tilde{H}} d_j^2(\hat{\mathcal{B}}_{LS}(\tilde{H}), \tilde{\Sigma}_{LS}(\tilde{H})). \quad (\text{A.14})$$

Using (A.9) 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 2 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 2 (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 (A.9) 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 $\tilde{H} \in \operatorname{argmin}_H \det \hat{\Sigma}_{LS}(H)$ which ends the proof. \square

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 $\|A\| = \sup_{\|u\|=1} \|Au\|$ where $u \in \mathbb{R}^q$ and $A \in \mathbb{R}^{p \times q}$. Sometimes we will also use the L_2 -norm $\|A\|_2 = (\sum_{i,j} |a_{ij}|^2)^{1/2}$. Since all norms on $\mathbb{R}^{p \times q}$ 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 $A \in \mathbb{R}^{p \times q}$.

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 $\gamma \in \mathbb{R}^p$. Furthermore it is excluded that there exists a $\mathcal{B} \in \mathbb{R}^{p \times q}$ 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 $\alpha \in \mathbb{R}^q$ 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 $\mathcal{B} \in \mathbb{R}^{p \times q}$ the $r_j := y_j - \mathcal{B}^t x_j$ are not lying on a $(q - 1)$ dimensional hyperplane, we have that

$$c_2(J) = \inf_{\mathcal{B} \in \mathbb{R}^{p \times q}} \lambda_{\min} \text{Cov}_0(\{r_j; j \in J\}) > 0$$

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

$$c = \min_J (\min(c_1(J), c_2(J))) > 0 \tag{A.15}$$

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

$$M = \sup_{H \in \mathcal{H}} \|\mathcal{B}_{LS}(H)^t\| < \infty \tag{A.16}$$

since no h points of $\{x_i; i = 1, \dots, n\}$ 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

$$\bar{M} = ((V h (\frac{h}{k(Z_n) + 1} c)^{1-q})^{1/2} + N_y) \frac{1}{\alpha_1 c} \quad (\text{A.17})$$

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 in [16, p. 244] and properties of norms it follows that

$$\begin{aligned} \det(\hat{\Sigma}_{LS}(H_1)) &\leq \lambda_{\max}(\text{cov}(\{r_j(\hat{\mathcal{B}}_{LS}(H_1)); j \in H_1\})^q \\ &\leq (\frac{1}{h} \sum_{j \in H_1} \lambda_{\max}(r_j(\hat{\mathcal{B}}_{LS}(H_1))r_j(\hat{\mathcal{B}}_{LS}(H_1))^t))^q \\ &= (\frac{1}{h} \sum_{j \in H_1} \|r_j(\hat{\mathcal{B}}_{LS}(H_1))\|^2)^q \\ &\leq (\frac{1}{h} \sum_{j \in H_1} (\|y_j\| + \|\hat{\mathcal{B}}_{LS}(H_1)^t x_j\|)^2)^q \\ &\leq (N_y + M N_x)^{2q} \\ &= V \end{aligned} \quad (\text{A.18})$$

where λ_{\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 again lemma 5.1 in [16] we obtain

$$\begin{aligned} \lambda_{\min}(\hat{\Sigma}_{LS}(H_2)) &= \lambda_{\min}(\text{cov}(\{y_j - \mathcal{B}_2^t x_j; j \in H_2\})) \\ &\geq \frac{k(Z_n) + 1}{h} \lambda_{\min}(\text{Cov}_0(\{y_j - \mathcal{B}_2^t x_j; j \in \bar{J}\})) \\ &\geq \frac{k(Z_n) + 1}{h} c_2(\bar{J}) \\ &\geq \frac{k(Z_n) + 1}{h} c \end{aligned} \quad (\text{A.19})$$

On the other hand,

$$\lambda_{\max}(\hat{\Sigma}_{LS}(H_2)) = \sup_{\|u\|=1} \frac{1}{h} \sum_{j \in H_2} u^t (y_j - \mathcal{B}_2^t x_j) (y_j - \mathcal{B}_2^t x_j)^t u. \quad (\text{A.20})$$

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

$$\|\mathcal{B}_2^t x_{j_0}\|^2 = \sum_{j=1}^q |\mathcal{B}_{2j} x_{j_0}|^2 \geq \sum_{j=1}^q \|\mathcal{B}_{2j}\|^2 c_1(\bar{J})^2 = (\|\mathcal{B}_2\|_2 c_1(\bar{J}))^2 \geq (\alpha_1 \|\mathcal{B}_2\| c_1(\bar{J}))^2$$

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 (A.20) that

$$\lambda_{\max}(\hat{\Sigma}_{LS}(H_2)) \geq \|y_{j_0} - \mathcal{B}_2^t x_{j_0}\|^2/h > (\alpha_1 \bar{M}c - N_y)^2/h. \quad (\text{A.21})$$

Combining (A.21) and (A.19) yields

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

by definition of \bar{M} . Together with (A.18) 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 $\tilde{J} \subset \{1, \dots, n\}$ 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 $\beta^t y_j - \gamma^t x_j = 0$ for all $j \in \tilde{J}$.

If $\beta \neq 0$ then there exists a $\mathcal{B} \in \mathbb{R}^{p+q}$ such that $\mathcal{B}\beta = \gamma$ which implies $\beta^t(y_j - \mathcal{B}^t x_j) = 0$ 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 $\lambda \rightarrow \infty$.

If $\beta = 0$ 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 \mathbb{R}^{p+q} we have that $\det \text{cov}(\{y_j - \hat{\mathcal{B}}_{LS}(H_2)^t x_j; j \in H_2\}) = 0$. But since $\gamma^t x = 0$ is a vertical hyperplane we have $\|\hat{\mathcal{B}}_{LS}(H_2)\| = \infty$ and it follows that $\|\hat{\mathcal{B}}_{MLTS}(Z'_n)\| = \infty$. \square

Proof of Corollary 1. The first part of the proof of Theorem 1 implies that $\varepsilon_n^*(\hat{\mathcal{B}}_{MLTS}^1, Z_n) \geq \min(n - h + 1, h - k(Z_n))/n$ and $\varepsilon_n^*(\hat{\mathcal{B}}_{MLTS}^0, Z_n) \geq \min(n - h + 1, h - k(Z_n))/n$. From the second part of the proof it follows that both $\|\hat{\mathcal{B}}_{MLTS}^1(Z'_n)\|$ and $\|\hat{\mathcal{B}}_{MLTS}^0(Z'_n)\|$ can be pulled towards ∞ when replacing more than $m = \min(n - h + 1, h - k(Z_n))$ data points by arbitrary points.

Proof of Corollary 2. 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 (A.19) 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 $k'(Z_n)$. \square

Proof of Corollary 3. Denote $Z_n = \{(1, y_i); 1 \leq i \leq n\}$ then clearly $\hat{\mu}_{MCD}(Y_n) = \hat{\mathcal{B}}_{MLTS}(Z_n)$ and $\hat{\Sigma}_{MCD}(Y_n) = \hat{\Sigma}_{MLTS}(Z_n)$. Moreover, the maximal number of points $k''(Y_n)$ from Y_n lying on a hyperplane of \mathbb{R}^q is equal to the maximal number of points from Z_n lying on a subspace of \mathbb{R}^{q+1} , hence $k(Z_n) = k''(Y_n)$. Therefore, the result immediately follows from the previous proof. \square

Proof of Theorem 2. Using properties of traces we obtain

$$\frac{1}{h} \sum_{j \in H_2} d_j^2(\hat{\mathcal{B}}_2, \hat{\Sigma}_2) = \frac{1}{h} \text{tr} \sum_{j \in H_2} r_j(\hat{\mathcal{B}}_2)^t \hat{\Sigma}_2^{-1} r_j(\hat{\mathcal{B}}_2) = \text{tr} \hat{\Sigma}_2^{-1} \hat{\Sigma}_2 = \text{tr} I_q = q \quad (\text{A.22})$$

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

$$c := \frac{1}{hq} \sum_{j \in H_2} d_j^2(\hat{\mathcal{B}}_1, \hat{\Sigma}_1) \leq \frac{1}{hq} \sum_{j \in H_1} d_j^2(\hat{\mathcal{B}}_1, \hat{\Sigma}_1) = 1, \quad (\text{A.23})$$

and also $c > 0$ since $\det(\hat{\Sigma}_2) > 0$. Combining (A.22) and (A.23) yields

$$\frac{1}{h} \sum_{j \in H_2} r_j(\hat{\mathcal{B}}_1)^t (c\hat{\Sigma}_1)^{-1} r_j(\hat{\mathcal{B}}_1) = \frac{1}{ch} \sum_{j \in H_2} d_j^2(\hat{\mathcal{B}}_1, \hat{\Sigma}_1) = \frac{cq}{c} = q. \quad (\text{A.24})$$

From Lemma 2 it follows that $\det(\hat{\Sigma}_2) \leq \det(c\hat{\Sigma}_1)$ and (A.23) implies $\det(c\hat{\Sigma}_1) \leq \det(\hat{\Sigma}_1)$, hence $\det(\hat{\Sigma}_2) \leq \det(\hat{\Sigma}_1)$. Moreover, from Lemma 2 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 $\hat{\Sigma}_2 = \hat{\Sigma}_1$. \square

Proof of Lemma 1. Clearly, we have that $\mathcal{E} \in D_H(\alpha)$. Note that

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

On the other hand, we have that

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

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

$$\frac{1}{1-\alpha} \int_{\mathcal{E}} (y - (\mathcal{B}_{\hat{A}}(H))^t x)^t (c \Sigma_{\hat{A}}(H))^{-1} (y - (\mathcal{B}_{\hat{A}}(H))^t x) dH = q \quad (\text{A.25})$$

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 (A.25) contradicts lemma 2 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. \square

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 $\mathcal{B}_{MLTS}(H) = 0$. Then we will have that $\Sigma_{MLTS}(H)$ is the MCD functional at the distribution of $y - \mathcal{B}_{MLTS}(H)^t x = y = \varepsilon$. Since the factor c_α makes the MCD Fisher-consistent at elliptical distributions (see[3,4] it will follow that $\Sigma_{MLTS}(H) = I_q$. Lemma 1 shows that \mathcal{B}_{MLTS} is the least squares fit based solely on the cylinder $\mathcal{C} = \{(x, y) \in \mathbb{R}^{p+q}; (y - \mathcal{B}_{MLTS}^t x)^t \Sigma_{MLTS}^{-1} (y - \mathcal{B}_{MLTS}^t x) \leq D_\alpha^2\}$. Therefore,

$$\int_{\mathcal{C}} x (y - \mathcal{B}_{MLTS}^t x)^t dH(x, y) = 0 \quad (\text{A.26})$$

Now suppose that $\mathcal{B}_{MLTS} \neq 0$. Let $\lambda_1, \dots, \lambda_q$ be the eigenvalues of Σ_{MLTS} and v_1, \dots, v_q the corresponding eigenvectors. There will be at least one $1 \leq j \leq q$ such that $\mathcal{B}_{MLTS} v_j \neq 0$. (Note that \mathcal{B}_{MLTS} is not necessarily of full rank.) Fix

this j . From (A.26) it follows that we should have

$$\int_{\mathcal{E}} v_j^t(\mathcal{B}_{LTS}^t x)(y - \mathcal{B}_{MLTS}^t x)^t v_j dF(y) dG(x) = 0$$

which can be rewritten as

$$\int_{\mathbb{R}^p} v_j^t(\mathcal{B}_{MLTS}^t x) I(x) dG(x) = 0 \quad (\text{A.27})$$

with

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

where $\mathcal{C}_x = \{y \in \mathbb{R}^q \mid (x, y) \in \mathcal{C}\}$. Fix x and set $d = (d_1, \dots, d_q)^t := \mathcal{B}_{MLTS}^t x$.

Since y is spherically symmetrically distributed, for computing $I(x)$ we may assume w.l.o.g. that $\Sigma_{MLTS} = \text{diag}(\lambda_1, \dots, \lambda_q)$ as well as $v_j = (1, 0, \dots, 0)$.

For every $d_1 - \sqrt{c\lambda_1} \leq y_1 \leq d_1 + \sqrt{c\lambda_1}$ denote

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

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

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

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

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

If $d_1 > 0$ we have $(d_1+t)^2 + y_2^2 + \dots + y_q^2 > (d_1-t)^2 + y_2^2 + \dots + y_q^2$ (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 $v_j^t(\mathcal{B}_{MLTS}^t x) = 0$, then $I(x) > 0$.

Also, $v_j^t(\mathcal{B}_{MLTS}^t x) = 0$ implies $I(x) = 0$. However, due to condition (24),

the latter event occurs with probability less than $1 - \alpha$. Therefore, we obtain

$\int_{\mathbb{R}^p} v_j^t \mathcal{B}_{MLTS}^t x I(x) dG(x) < 0$ which contradicts (A.27), so we conclude that $\mathcal{B}_{MLTS} = 0$. \square

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 $\mathcal{B}_\varepsilon := \mathcal{B}_{MLTS}(H_\varepsilon)$ and $\Sigma_\varepsilon := \Sigma_{MLTS}(H_\varepsilon)$. Then (20) results in

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

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

$$\begin{aligned} IF(z_0; \mathcal{B}_{MLTS}, H_0) &= \left(\int_{\hat{A}} xx^t dH_0(z) \right)^{-1} \frac{\partial}{\partial \varepsilon} \int_{\hat{A}_\varepsilon} xy^t dH_\varepsilon(z) \Big|_{\varepsilon=0} \\ &\quad + \frac{\partial}{\partial \varepsilon} \left[\left(\int_{\hat{A}_\varepsilon} xx^t dH_\varepsilon(z) \right)^{-1} \right] \Big|_{\varepsilon=0} \int_{\hat{A}} xy^t dH_0(z) \end{aligned}$$

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

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

by symmetry of F and

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

Therefore, we obtain

$$\begin{aligned} IF(z_0; \mathcal{B}_{MLTS}, H_0) &= \frac{E_G[xx^t]^{-1}}{1 - \alpha} \frac{\partial}{\partial \varepsilon} \int_{\hat{A}_\varepsilon} xy^t dH_\varepsilon(z) \Big|_{\varepsilon=0} \\ &= \frac{E_G[xx^t]^{-1}}{1 - \alpha} \frac{\partial}{\partial \varepsilon} \left((1 - \varepsilon) \int_{\hat{A}_\varepsilon} xy^t dH_0(z) + \varepsilon x_0 y_0^t I(z_0 \in \hat{A}_\varepsilon) \right) \Big|_{\varepsilon=0} \\ &= \frac{E_G[xx^t]^{-1}}{1 - \alpha} \left(x_0 y_0^t I(\|y_0\|^2 \leq q_\alpha) + \frac{\partial}{\partial \varepsilon} \int_{\hat{A}_\varepsilon} xy^t dH_0(z) \right). \quad (\text{A.28}) \end{aligned}$$

Similarly to Proposition 1 in [4], it can be shown that Lemma 1 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 $\hat{A}_\varepsilon = \{(x, y) \in \mathbb{R}^{p+q}; d_\varepsilon^2(x, y) \leq q_\alpha(\varepsilon)\}$ 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 $\mathcal{E}_{\varepsilon, x} := \{y \in \mathbb{R}^q; d_\varepsilon^2(x, y) \leq q_\alpha(\varepsilon)\}$. Then it follows that

$$\int_{\hat{A}_\varepsilon} xy^t dH_0(z) = \int_{\mathbb{R}^p} \int_{\mathcal{E}_{\varepsilon, x}} xy^t dF(y) dG(x) = \int_{\mathbb{R}^p} x \left(\int_{\mathcal{E}_{\varepsilon, x}} y g(y^t y) dy \right)^t dG(x). \quad (\text{A.29})$$

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 $\theta = (\theta_1, \dots, \theta_{q-1}) \in \Theta = [0, \pi[\times \dots \times [0, \pi[\times [0, 2\pi[$, 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

$$\frac{\partial}{\partial \varepsilon} I(\varepsilon) \Big|_{\varepsilon=0} = \int_{\|v\|^2 \leq q_\alpha} \frac{\partial}{\partial \varepsilon} \left((\Sigma_\varepsilon^{1/2} v + \mathcal{B}_\varepsilon^t x) g((\Sigma_\varepsilon^{1/2} v + \mathcal{B}_\varepsilon^t x)^t (\Sigma_\varepsilon^{1/2} v + \mathcal{B}_\varepsilon^t x)) \right) \Big|_{\varepsilon=0} dv \quad (\text{A.30})$$

The derivative on the right hand side becomes

$$\begin{aligned} & \frac{\partial}{\partial \varepsilon} \{ (\Sigma_\varepsilon^{1/2} v + \mathcal{B}_\varepsilon^t x) g((\Sigma_\varepsilon^{1/2} v + \mathcal{B}_\varepsilon^t x)^t (\Sigma_\varepsilon^{1/2} v + \mathcal{B}_\varepsilon^t x)) \} \Big|_{\varepsilon=0} = \{ IF(z_0; \Sigma_{MLTS}^{1/2}, H_0) v \\ & + IF(z_0; \mathcal{B}_{MLTS}, H_0)^t x \} g(v^t v) + 2 v g'(v^t v) \{ (v^t IF(z_0; \Sigma_{MLTS}^{1/2}, H_0) v + v^t IF(z_0; \mathcal{B}_{MLTS}, H_0)^t x) \} \end{aligned} \quad (\text{A.31})$$

Since $\int_{\|v\|^2 \leq q_\alpha} v g(v^t v) dv$ and $\int_{\|v\|^2 \leq q_\alpha} v g'(v^t v) v^t IF(z_0; \Sigma_{MLTS}^{1/2}, H_0) v dv$ are zero due to symmetry of F , the terms in (A.31) including $IF(z_0; \Sigma_{MLTS}^{1/2}, H_0)$ give a zero contribution to the integral in (A.30). It follows that

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

where $c_2 = \int_{\|v\|^2 \leq q_\alpha} g'(v^t v) v_1^2 dv$ can be rewritten in the form given in Theorem 4 by using polar coordinates. From (A.29) we now obtain that

$$\frac{\partial}{\partial \varepsilon} \int_{\hat{A}_\varepsilon} xy^t dH_0(z)|_{\varepsilon=0} = [(1 - \alpha) + 2c_2] E_G[xx^t] IF(z_0; \mathcal{B}_{MLTS}, H_0). \quad (\text{A.32})$$

Substituting (A.32) in (A.28) yields

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

which results in

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

References

- [1] 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.
- [2] Bilodeau, M. and Duchesne P. (2000), Robust Estimation of the SUR Model, *Canadian Journal of Statistics*, **28**, 277–288.
- [3] Butler, R.W., Davies, P.L., and Jhun, M. (1993), Asymptotics for the Minimum Covariance Determinant Estimator, *The Annals of Statistics*, **21**, 1385–1400.
- [4] Croux, C., and Haesbroeck, G. (1999), Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator, *Journal of Multivariate Analysis*, **71**, 161–190.
- [5] Croux, C., Rousseeuw, P.J., and Hössjer, O. (1994), Generalized S-estimators, *Journal of the American Statistical Association*, **89**, 1271–1281.

- [6] 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.
- [7] García Ben, M., Martínez, E., and Yohai, V.J. (2005), Robust Estimation for the Multivariate Linear Model Based on a τ -scale,” *Journal of Multivariate Analysis*, to appear.
- [8] Grübel, R. (1988), A Minimal Characterization of the Covariance Matrix, *Metrika*, **35**, 49–52.
- [9] 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.
- [10] Hawkins, D.M. and Olive, D. (2002), Inconsistency of Resampling Algorithms for High-Breakdown Regression Estimators and a New algorithm, *Journal of the American Statistical Association*, **97**, 136–159.
- [11] Hössjer, O. (1994), Rank-Based Estimates in the Linear Model With High Breakdown Point,” *Journal of the American Statistical Association*, **89**, 149–158.
- [12] Johnson, N.L., and Kotz, S. (1972), *Distributions in Statistics: Continuous Multivariate distributions*, John Wiley and Sons, New York.
- [13] Koenker, R., and Portnoy, S. (1990), M Estimation of Multivariate Regressions, *Journal of the American Statistical Association*, **85**, 1060–1068.
- [14] Kung, K.-M. (2005), Multivariate Least-Trimmed Squares Regression Estimator, *Computational Statistics and Data Analysis*, **48**, 307–316.
- [15] Lopuhaä, H.P. (1989), On the Relation Between S-Estimators and M-Estimators of Multivariate Location and Covariance, *The Annals of Statistics*, **17**, 1662–1683.

- [16] 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.
- [17] Marrona, R.A., and Yohai, V.J. (1997), Robust Estimation in Simultaneous Equations Models, *Journal of Statistical Planning and Inference*, **57**, 233–244.
- [18] Ollila, E., Oja, H., and Hettmansperger, T.P. (2002), Estimates of Regression Coefficients Based on the Sign Covariance Matrix, *Journal of the Royal Statistical Society, Ser. B*, **64**, 447–466.
- [19] Ollila, E., Oja, H., and Koivunen, V. (2003), Estimates of Regression Coefficients Based on Lift Rank Covariance Matrix, *Journal of the American Statistical Association*, **98**, 90–98.
- [20] Rousseeuw, P.J. (1984), Least Median of Squares Regression, *Journal of the American Statistical Association*, **79**, 871–880.
- [21] Rousseeuw, P.J., and Leroy, A.M. (1987), *Robust regression and outlier detection*, Wiley-Interscience, New York.
- [22] Rousseeuw, P.J., Van Aelst, S., Van Driessen, K., and Agull, J. (2004), Robust Multivariate Regression, *Technometrics*, **46**, 293–305.
- [23] Rousseeuw, P.J., and Van Driessen K. (1999), A Fast Algorithm for the Minimum Covariance Determinant Estimator, *Technometrics*, **41**, 212–223.
- [24] Ruppert, D. (1992), Computing S-estimators for Regression and Multivariate Location/Dispersion,” *Journal of Computational and Graphical Statistics*, **1**, 253–270.
- [25] Van Aelst, S., and Willems, G. (2005), Multivariate Regression S-Estimators for Robust Estimation and Inference, *Statistica Sinica*, **15**, 981–1001.

- [26] Woodruff D.L. and Rocke, D.L. (1994), Computable Robust Estimation of Multivariate Location and Shape in High Dimension Using Compound Estimators, *Journal of the American Statistical Association*, **89**, 888–896.