The multivariate least-trimmed squares estimator
Jose Agulló\textsuperscript{a}, Christophe Croux\textsuperscript{b}, Stefan Van Aelst\textsuperscript{c},\textsuperscript{*}

\textsuperscript{a}Departamento de Fundamentos del Análisis Económico, University of Alicante, E-03080, Alicante, Spain
\textsuperscript{b}Faculty of Economics and Applied Economics, ORSTAT Research Center, KU Leuven, Naamsestraat 69, B-3000 Leuven, Belgium
\textsuperscript{c}Department of Applied Mathematics and Computer Science, Ghent University (UGENT), Krijgslaan 281 S9, B-9000 Gent, Belgium

Received 10 March 2006
Available online 17 August 2006

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.

© 2006 Elsevier Inc. All rights reserved.

AMS 2000 subject classification: 62F35; 62H12

Keywords: Multivariate regression; Breakdown point; Influence function; Minimum covariance determinant estimator

1. Introduction
Consider the multivariate regression model
\begin{equation}
y_i = B^t x_i + \epsilon_i, \tag{1}
\end{equation}
where \( i = 1, \ldots, n \) with \( x_i = (x_{i1}, \ldots, x_{ip})^t \in \mathbb{R}^p \) and \( y_i = (y_{i1}, \ldots, y_{iq})^t \in \mathbb{R}^q \). The matrix \( B \in \mathbb{R}^{p \times q} \) contains the regression coefficients. The error terms \( \epsilon_1, \ldots, \epsilon_n \) are i.i.d. with zero center and as covariance a positive definite and symmetric matrix \( \Sigma \) 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 \( Z_n = \{(x_i, y_i); i = 1, \ldots, n\} \) and write \( X = (x_1, \ldots, x_n)^t \) for the design matrix and \( Y = (y_1, \ldots, y_n)^t \) for the matrix of responses. The classical estimator for \( B \) is the least-squares (LS) estimator \( \hat{B}_{LS} \) which is given by

\[
\hat{B}_{LS} = (X^t X)^{-1} X^t Y
\]  

(2)

while \( \Sigma \) is unbiasedly estimated by

\[
\hat{\Sigma}_{LS} = \frac{1}{n - p} (Y - X\hat{B}_{LS})^t (Y - X\hat{B}_{LS}).
\]  

(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 [14] 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 (MCD) 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 MLTS 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 \( X = (1, \ldots, 1)^t \in \mathbb{R}^n \). 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 \( \beta \) 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 \( \tau \)-estimators [7].
We thus use the 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 data set \( Z_n = \{(x_i, y_i); i = 1, \ldots, n\} \subset \mathbb{R}^{p+q} \) and for any \( B \in \mathbb{R}^{p \times q} \) denote \( r_i(B) = y_i - B^T x_i \) 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 \( \hat{B}_{LS}(H) \) the least-squares fit based solely on the observations \( \{(x_j, y_j); j \in H\} \). Furthermore, for any \( H \in \mathcal{H} \) and \( B \in \mathbb{R}^{p \times q} \) denote

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

with \( \bar{r}_H(B) := \frac{1}{h} \sum_{j \in H} r_j(B) \), the covariance matrix of the residuals with respect to the fit \( B \), belonging to the subset \( H \). Then the MLTS estimator is defined as follows:

**Definition 1.** With the notations above, the MLTS estimator is defined as

\[
\hat{B}_{MLTS}(Z_n) = \hat{B}_{LS}(\hat{H}) \quad \text{where} \quad \hat{H} \in \arg\min_{H \in \mathcal{H}} \det \hat{\Sigma}_{LS}(H) \quad (5)
\]

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

\[
\hat{\Sigma}_{MLTS}(Z_n) = c_2 \hat{\Sigma}_{LS}(\hat{H}), \quad (6)
\]

where \( c_2 \) is a consistency factor.

Note that if the minimization problem has more than one solution, in which case we look at \( \arg\min_H \det \hat{\Sigma}_{LS}(H) \) as a set, we arbitrarily select one of these solutions to determine the MLTS estimator. In Section 5 a consistency factor \( c_2 \) 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 \( \Sigma \) 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 data set \( Z_n = \{(x_i, y_i); i = 1, \ldots, n\} \subset \mathbb{R}^{p+q} \) are lying in the same subspace of \( \mathbb{R}^{p+q} \). Formally, this means that for all \( \beta, \gamma \in \mathbb{R}^p, \gamma \in \mathbb{R}^q \), it holds that

\[
\#\{(x_i, y_i) | \beta^T x_i + \gamma^T y_i = 0\} < h \quad \text{unless if} \ \beta \ \text{and} \ \gamma \ \text{are both zero vectors.}
\]

For data sets 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 \( \hat{H} \) minimizing the determinant of the MCD scatter matrix estimate computed from its residuals. For any \( B \in \mathbb{R}^{p \times q} \), let us denote \( \text{MCD}_q(B) \) the MCD-scatter matrix based on the residuals from \( B \). Formally,

\[
\text{MCD}_q(B) = \text{Cov}_0(\hat{H}, B) = \frac{1}{h} \sum_{j \in \hat{H}} r_j(B)r_j(B)^T
\]

with \( \hat{H} \in \arg\min_{H \in \mathcal{H}} \det \text{Cov}_0(H, B) \). 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 data sets satisfying (7), we have that
\[
\arg\min_{\mathcal{B}} \det MCD_q(\mathcal{B}) = \left\{ \hat{\mathcal{B}}_{\text{LS}}(\hat{\mathcal{H}}) \mid \hat{\mathcal{H}} \in \arg\min_{\mathcal{H} \in \mathcal{H}} \hat{\Sigma}_{\text{LS}}(\mathcal{H}) \right\}. \tag{8}
\]

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 distribution, we can rewrite (8) as
\[
\hat{\mathcal{B}}_{\text{MLTS}}(Z_n) = \arg\min_{\mathcal{B}} \det MCD_q(\mathcal{B}). \tag{9}
\]

For the residual scatter estimator we have
\[
\hat{\Sigma}_{\text{MLTS}}(Z_n) = c_q MCD_q(\hat{\mathcal{B}}_{\text{MLTS}}(Z_n)). \tag{10}
\]

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 \( \Sigma \in \text{PDS}(q) \), the class of positive definite and symmetric matrices of size \( q \), we define the squared distances (for the \( \Sigma \) metric) of the residuals w.r.t. \( \mathcal{B} \) as
\[
d^2_i(\mathcal{B}, \Sigma) := r_i(\mathcal{B})' \Sigma^{-1} r_i(\mathcal{B}).
\]
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
\[
\arg\min_{\mathcal{B}, \Sigma; |\Sigma| = 1} \sum_{j=1}^h d^2_{j:n}(\mathcal{B}, \Sigma), \tag{11}
\]
where the minimum is over all \( \mathcal{B} \in \mathbb{R}^{p \times q} \) and \( \Sigma \in \text{PDS}(q) \) with \( \det \Sigma = 1 \) (denoted as \( |\Sigma| = 1 \)). Then for data sets satisfying (7) it holds that
\[
\left\{ \hat{\mathcal{B}} \mid (\hat{\mathcal{B}}, \hat{\Sigma}) \in \arg\min_{\mathcal{B}, \Sigma; |\Sigma| = 1} \sum_{j=1}^h d^2_{j:n}(\mathcal{B}, \Sigma) \right\} = \left\{ \hat{\mathcal{B}}_{\text{LS}}(\hat{\mathcal{H}}) \mid \hat{\mathcal{H}} \in \arg\min_{\mathcal{H} \in \mathcal{H}} \hat{\Sigma}_{\text{LS}}(\mathcal{H}) \right\}. \tag{12}
\]

Proposition 2 shows that any \( \hat{\mathcal{B}} \) 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
\[
\hat{\mathcal{B}}_{\text{MLTS}}(Z_n) = \arg\min_{\mathcal{B}, \Sigma; |\Sigma| = 1} \sum_{j=1}^h d^2_{j:n}(\mathcal{B}, \Sigma), \tag{13}
\]
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 \( \arg\min_{\mathcal{B}} \sum_{j=1}^h r^2_{j:n}(\mathcal{B}) \), 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 (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 BDP. The finite-sample BDP $\varepsilon_n^*$ 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 data set $Z_n \subset \mathbb{R}^{p+q}$ 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 data set $Z_n \subset \mathbb{R}^{p+q}$ satisfying (7) with $q > 1$ it holds that

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

(14)

It follows that if we take $h = \gamma n$ for some fraction $0 < \gamma \leq 1$ then the corresponding BDP equals $\varepsilon_n^*(\hat{B}_{MLTS}, Z_n) = \min(1 - \gamma + 1/n, \gamma - k(Z_n)/n)$. If the data set $Z_n$ 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 $\varepsilon_n^*(\hat{B}_{MLTS}, Z_n) = \min(n - h + 1, h - p - q + 1)/n$ almost surely. Then for $h = \gamma n$ the BDP of the MLTS tends to $\min(1 - \gamma, \gamma)$. It follows that for data with $k(Z_n) = p + q - 1$ any choice $\left[ (n + p + q)/2 \right] \leq h \leq \left[ (n + p + q + 1)/2 \right]$ yields the maximal BDP, which is $\left[ (n - p - q)/2 \right] + 1)/n \approx 50\%$.

**Remark 1.** For a regression model with intercept we can explicitly write $x_i = (u'_i, 1)'$ with $u_i = (x_{i1}, \ldots, x_{ip-1})'$ in (1). In this case, the last row of $B$ is the intercept vector $(B^0)'$ and the matrix formed by the $p - 1$ first rows of $B$ is the slope matrix $B^1$. The previous result holds for both the slope matrix and intercept vector.

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

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

(15)

**Remark 2.** In the case $q = 1$, the proof of Theorem 1 becomes much easier and yields the following result for the BDP of the LTS estimator.

**Corollary 2.** Denote $k'(Z_n)$ the maximal number of $x_i \in \{x_i; i = 1, \ldots, n\}$ lying on a subspace of $\mathbb{R}^p$. Then for any data set $Z_n \subset \mathbb{R}^{p+1}$ with $k'(Z_n) < h$ it holds that

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

(16)

If $Z_n$ comes from a continuous distribution $H$ then almost surely $k'(Z_n) = p - 1$ yielding $\varepsilon_n^*(\hat{B}_{LTS}, Z_n) = \min(n - h + 1, h - p + 1)/n$, as was already obtained in [11]. In this case any $\left[ (n + p)/2 \right] \leq h \leq \left[ (n + p + 1)/2 \right]$ gives the maximal BDP.
Remark 3. In the case \( p = 1 \) and \( x_i = 1 \) Theorem 1 gives the following result for the BDP of the MCD estimator of multivariate location and scatter.

Corollary 3. Consider \( Y_n = \{y_1, \ldots, y_n\} \subset \mathbb{R}^q \) 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 data set \( Y_n \subset \mathbb{R}^q \) with \( k''(Y_n) < h \) it holds that

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

If \( Y_n \) comes from a continuous distribution \( F \) then almost surely \( k'(Z_n) = q - 1 \) 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 BDP.

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 \( \hat{B}_1 := \hat{B}_{LS}(H_1) \) 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 \( d_{1:n}(\hat{B}_1, \hat{\Sigma}_1) \leq \cdots \leq d_{h:n}(\hat{B}_1, \hat{\Sigma}_1) \). For \( \hat{B}_2 := \hat{B}_{LS}(H_2) \) 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 \( \hat{B}_2 = \hat{B}_1 \) 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 \( \{1, \ldots, n\} \) and compute the corresponding least-squares estimates \( \hat{B}_m := \hat{B}_{LS}(J_m) \) 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 \( d_i(\hat{B}_m, \hat{\Sigma}_m) \) 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 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 data sets the algorithm can be speeded 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 ML TS 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 ML TS estimator and define

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

To define the ML TS estimator at the distribution $K$ we require that

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

For each $A \in D_K(\alpha)$, the least-squares solution over the set $A$ is then given by

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

and

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

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

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

The constant $c_\alpha$ can be chosen such that consistency will be obtained at the specified model. If the distribution $K$ is not continuous, then the definition of $D_K(\alpha)$ can be modified as in [4] to ensure that the set $D_K(\alpha)$ is non-empty.

Now consider the multivariate regression model

$$y = B'x + \varepsilon, \quad (23)$$

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

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

where $\Sigma \in \text{PDS}(q)$. Furthermore, the function $g$ is assumed to have a strictly negative derivative $g'$ such that $F_\Sigma$ is a unimodal elliptically symmetric distribution around the origin. Note that we do not need that $F_\Sigma$ has finite second moments, but if second moments exist, then $\Sigma$ 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 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 $1 - \alpha$. Note that condition (24) implies condition (19) because $\gamma$ can be put equal to zero. We first give the following proposition which says that the ML TS 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 $d^2(x, y) = (y - B_{\hat{A}}(H)^T x)^T (\Sigma_{\hat{A}}(H))^{-1} (y - B_{\hat{A}}(H)^T x)$. Define the cylinder $E = \{(x, y) \in \mathbb{R}^{p+q}; d^2(x, y) \leq D_2^2\}$ where $D_2^2$ is chosen such that $P_H(E) = 1 - \alpha$. Then it holds that

$$B_E(H) = B_{\hat{A}}(H) \quad \text{and} \quad \Sigma_E(H) = \Sigma_{\hat{A}}(H).$$

We now show that the functionals $B_{\text{MLTS}}(H)$ and $\Sigma_{\text{MLTS}}(H)$ defined by (22) for some well chosen constant $c_\alpha$ are Fisher-consistent for the parameters $B$ and $\Sigma$.

**Theorem 3.** Denote

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

where $F_0 = F_{I_q}$ is the central error distribution and $q_x = F_{*}^{-1}(1 - \alpha)$ with $F_{*}(t) = P_{F_0}(U^T U \leq t)$. Then the functionals $B_{\text{MLTS}}$ and $\Sigma_{\text{MLTS}}$ are Fisher-consistent estimators for the parameters $B$ and $\Sigma$ at the model distribution $H$:

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

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 $c_\alpha$ reduces to $c_\alpha = (1 - \alpha)/F_{I_{q+2}}(q_x)$ with $q_x = \chi_{q,1-\alpha}^2$, the upper $\alpha$ percent point of the $\chi^2$ distribution with $q$ degrees of freedom and $F_{I_{q+2}}$ the cumulative distribution function of the $\chi^2$ 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 $\Delta_z$ and consider the contaminated distribution $H_{\varepsilon,z} = (1 - \varepsilon)H + \varepsilon \Delta_z$ then the influence function is given by

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

\[
IF(z; \mathcal{B}_{MLTS}, H_0) = E_{H_0}[x x']^{-1} \frac{xy}{-2c_2} I(\|y\|^2 \leq q_x),
\]

(25)

where \( c_2 \) is given by

\[
c_2 = \frac{\pi^{q/2}}{\Gamma(q/2 + 1)} \int_0^{\sqrt{q_x}} r^{q+1} g'(r^2) dr.
\]

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 ML TS 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

\[
IF(y, T_q, F_0) = y - \frac{1}{-2c_2} x I(\|y\|^2 \leq q_x).
\]

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

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

(26)

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

\[
IF(z; \mathcal{B}_{MLTS}, H_0) = E_{H_0}[x x']^{-1} \frac{xy I(\|y\|^2 \leq q_x)}{1 - \alpha - 2\sqrt{q_x} f(\sqrt{q_x})}
\]

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 \( x = (u', 1)' \) with \( u = (x_1, \ldots, x_{p-1})' \) in (23). In this case, the last row of \( \mathcal{B} \) is the intercept vector \((\mathcal{B}^0)'\) 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 \( \Sigma_u = E[(u - \mu_u)(u - \mu_u)'] \), it then follows immediately from (25) and (26) that

\[
IF(z; \mathcal{B}^1_{MLTS}, H_0) = \Sigma_u^{-1}(u - \mu_u) \frac{y'}{-2c_2} I(\|y\|^2 \leq q_x)
\]

\[
= \Sigma_u^{-1}(u - \mu_u)IF(y; T_q, F_0)'
\]

\[
IF(z; \mathcal{B}^0_{MLTS}, H_0) = (y - y(u - \mu_u)'\Sigma_u^{-1} \mu_u) \frac{I(\|y\|^2 \leq q_x)}{-2c_2}
\]

\[
= IF(y; T_q, F_0) - IF(z; \mathcal{B}^1_{MLTS}, H_0) \mu_u.
\]
The asymptotic variance–covariance matrix of \( \mathcal{B}_{\text{MLTS}} \) can now be computed by means of \( \text{ASV}(\mathcal{B}_{\text{MLTS}}, H_0) = E_H[IF(z; \mathcal{B}_{\text{MLTS}}, H_0) \otimes IF(z; \mathcal{B}_{\text{MLTS}}, H_0)^t] \) (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 \((d_1d_3 \times d_2d_4)\) matrix with \( d_1d_2 \) 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 \( \Sigma_x := E_{H_0}[xx'] \), then expression (26) implies that
\[
\text{ASV}(\mathcal{B}_{\text{MLTS}}, H_0) = D_{p,q} (\text{diag}(\text{ASV}(T_q, F)) \otimes \Sigma_x^{-1})
\]  
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 \( \Delta_{ml} \) 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 \( (\mathcal{B}_{\text{MLTS}})_{ij} \) is given by \( \Delta_{ji} \Sigma_x^{-1} \text{ASV}((T_q)_j, F) \) which implies that the asymptotic variance of \( (\mathcal{B}_{\text{MLTS}})_{ij} \) equals
\[
\text{ASV}((\mathcal{B}_{\text{MLTS}})_{ij}, H_0) = E_H[IF^2(z; (\mathcal{B}_{\text{MLTS}})_{ij}, H_0)] = (\Sigma_x^{-1})_{ii} \text{ASV}((T_q)_j, F).
\]
For \( i \neq i' \) we obtain the asymptotic covariances
\[
\text{ASC}((\mathcal{B}_{\text{MLTS}})_{ij}, (\mathcal{B}_{\text{MLTS}})_{i',j}, H_0) = E_H[IF(z; (\mathcal{B}_{\text{MLTS}})_{ij}, H_0)IF(z; (\mathcal{B}_{\text{MLTS}})_{i',j}, H_0)] = (\Sigma_x^{-1})_{ii'} \text{ASV}((T_q)_j, F)
\]
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 \( \text{ASV}((\mathcal{B}_{\text{MLTS}})_{ij}, H_0) = \text{ASV}((T_q)_j, F_0) \) for all \( 1 \leq i \leq p \) and \( 1 \leq j \leq q \). The limit case \( x = 0 \) yields the asymptotic variance of the least-squares estimator \( \text{ASV}((\mathcal{B}_{\text{LS}})_{ij}, H_0) = \text{ASV}(M_j, F) \) 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
\[
\text{ARE}((\mathcal{B}_{\text{MLTS}})_{ij}, H_0) = \frac{\text{ASV}((\mathcal{B}_{\text{LS}})_{ij}, H_0)}{\text{ASV}((\mathcal{B}_{\text{MLTS}})_{ij}, H_0)} = \frac{\text{ASV}(M_j, F_0)}{\text{ASV}((T_q)_j, F_0)} = \text{ARE}((T_q)_j, F_0)
\]
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

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 \( \hat{\mathcal{B}}_{\text{MLTS}} \) and \( \hat{\Sigma}_{\text{MLTS}} \) denote the initial MLTS estimates. Then the one-step reweighted MLTS estimates (RMLTS) are defined as
\[
\hat{\mathcal{B}}_{\text{RMLTS}} := \hat{\mathcal{B}}_{\text{LS}}(J) \quad \text{and} \quad \hat{\Sigma}_{\text{RMLTS}} := c_\delta \text{cov}(J, \hat{\mathcal{B}}_{\text{LS}}(J)),
\]
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 $q$

<table>
<thead>
<tr>
<th>$q$</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>10</th>
<th>30</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>10</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLTS</td>
<td>0.25</td>
<td>0.403</td>
<td>0.466</td>
<td>0.531</td>
<td>0.664</td>
<td>0.941</td>
<td>0.954</td>
<td>0.965</td>
<td>0.974</td>
<td>0.982</td>
</tr>
<tr>
<td>RMLTS</td>
<td>0.5</td>
<td>0.153</td>
<td>0.204</td>
<td>0.262</td>
<td>0.327</td>
<td>0.934</td>
<td>0.951</td>
<td>0.963</td>
<td>0.973</td>
<td>0.982</td>
</tr>
</tbody>
</table>

where $J = \{ j : d^2(\hat{B}_{MLTS}, \hat{S}_{MLTS}) \leq q_\delta \}$. Here $\delta$ 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 $q_\delta = \chi^2_{q,1-\delta}$ the corresponding quantile of the $\chi^2$-distribution with $q$ degrees of freedom. In the case of multivariate normal errors we have $c_\delta = (1 - \delta)/F_{\chi^2_{q,1-\delta}}(q_\delta)$.

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

$$B_{RMLTS}(K) = \left( \int f_j xx^t dK(x, y) \right)^{-1} \int f_j xy^t dK(x, y),$$

$$\Sigma_{RMLTS}(K) = c_\delta \int f_j (y - B_{RMLTS}(K)tx)(y - B_{RMLTS}(K)tx)^t dK(x, y),$$

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

$$IF(z, \hat{B}_{RMLTS}, H_0) = \left( 1 + \frac{2d_2}{1 - \delta} \right) IF(z, \hat{B}_{MLTS}, H_0)$$

$$+ \frac{EH_0[xx^t]^{-1}}{1 - \delta} xy^t I(||y||^2 \leq q_\delta),$$

where the constant $d_2$ is the same as $c_2$ in Theorem 4 but with $x$ replaced by $\delta$. 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 $B_{RMLTS}$ can be rewritten as

$$IF(z; B_{RMLTS}, H_0) = EH_0[xx^t]^{-1} x IF(y; T_q^{1}, F_0).$$

**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 \( B_{RMLTS} \) equals

\[
ASV(B_{RMLTS}, H_0) = D_{p,q} (\text{diag}(ASV(T_q^1, F)) \otimes \Sigma_{\chi}^{-1}).
\]  

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

\[
ASV((B_{MLTS})_{ij}, H_0) = (\Sigma_{\chi}^{-1})_{ii} ASV((T_q^1)_j, F),
\]

\[
ASC((B_{MLTS})_{ij}, (B_{MLTS})_{i'j}, H_0) = (\Sigma_{\chi}^{-1})_{ii'} ASV((T_q^1)_j, F) \quad \text{for } i \neq i'
\]

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

\[
ARE((B_{RMLTS})_{ij}, H_0) = ARE((T_q^1)_j, F_0)
\]

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 data sets of size \( n = 100, 300 \) and 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 \( V \sim \chi^2_{p-1} \). (See e.g. [12, p. 134].)

W.l.o.g. we took \( B = 0 \) 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

\[
\text{MSE}(T) = \text{average}(T^{(l)} - \theta)^2, \quad l = 1, \ldots, m,
\]
Table 2
MSE of the slope at normal (NOR) or exponential (EXP) carrier distributions and normal error distribution

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Method</th>
<th>( n = 100 )</th>
<th>( n = 300 )</th>
<th>( n = 500 )</th>
<th>( n = \infty )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NOR</td>
<td>EXP</td>
<td>NOR</td>
<td>EXP</td>
</tr>
<tr>
<td>( p = 3 )</td>
<td>M</td>
<td>1.10</td>
<td>1.19</td>
<td>1.09</td>
<td>1.08</td>
</tr>
<tr>
<td></td>
<td>MLTS</td>
<td>4.09</td>
<td>4.99</td>
<td>4.42</td>
<td>4.77</td>
</tr>
<tr>
<td></td>
<td>RMLTS</td>
<td>1.79</td>
<td>2.25</td>
<td>1.21</td>
<td>1.30</td>
</tr>
<tr>
<td></td>
<td>( p = 3 )</td>
<td>S</td>
<td>1.52</td>
<td>1.70</td>
<td>1.42</td>
</tr>
<tr>
<td></td>
<td>MCD</td>
<td>7.94</td>
<td>21.25</td>
<td>8.48</td>
<td>20.74</td>
</tr>
<tr>
<td></td>
<td>LRMCD</td>
<td>1.56</td>
<td>2.44</td>
<td>1.12</td>
<td>1.22</td>
</tr>
<tr>
<td>( q = 3 )</td>
<td>M</td>
<td>1.21</td>
<td>1.28</td>
<td>1.10</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td>MLTS</td>
<td>3.17</td>
<td>3.99</td>
<td>3.52</td>
<td>3.94</td>
</tr>
<tr>
<td></td>
<td>RMLTS</td>
<td>2.35</td>
<td>3.00</td>
<td>1.31</td>
<td>1.42</td>
</tr>
<tr>
<td></td>
<td>( q = 3 )</td>
<td>S</td>
<td>1.33</td>
<td>1.52</td>
<td>1.25</td>
</tr>
<tr>
<td></td>
<td>MCD</td>
<td>4.51</td>
<td>6.50</td>
<td>4.33</td>
<td>6.21</td>
</tr>
<tr>
<td></td>
<td>LRMCD</td>
<td>3.33</td>
<td>4.42</td>
<td>1.77</td>
<td>1.67</td>
</tr>
</tbody>
</table>

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

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

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% BDP. Results for the multivariate M-estimator proposed in [14] are included as well. The M-estimator uses the Huber 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 uncorrelated. 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 Fig. 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. Fig. 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 data sets. To generate contaminated data sets we started from the uncontaminated
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%.

Data sets as before and then we replaced 20% of the data with observations for which the \( p-1 \) independent variables were generated according to \( N(\lambda \sqrt{\chi^2_{p-1,0.99}}, 1.5) \) and the \( q \) dependent variables were generated from \( N(\kappa \sqrt{\chi^2_{q,0.99}}, 1.5) \). Both \( \lambda \) and \( \kappa \) took values in \( \{0, 0.5, 1, 1.5, 2, 3, 4, 5\} \). 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 \( \lambda \) and \( \kappa \) produce extreme outliers while small values produce intermediate outliers.

In Fig. 2 we show for each value of \( \lambda \) the maximal value of MSE obtained over all possible values of \( \kappa \). We show results for data sets 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 Fig. 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 Fig. 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 LRMCMD 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 MLTS 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 BDP 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 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 BDP 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 MLTS regression estimator [13]. 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 BDP 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.
Appendix A

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 \( E_K[xx'] \) is a strictly positive definite matrix. Define \( B_{LS}(K) = E_K[xx']^{-1}E_K[xy'] \) and \( \Sigma_{LS}(K) = \text{Cov}_0(\varepsilon) := E_K[\varepsilon\varepsilon'] \) where \( \varepsilon := y - (B_{LS}(K))^\dagger x \). Then among all pairs \((b, \Delta)\) with \( b \in \mathbb{R}^{p \times q} \) and \( \Delta \) a positive definite and symmetric matrix of size \( q \) such that

\[
E_K[(y - b'x)\Delta^{-1}(y - b'x)] = q,
\]

(A.1)

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

Note that if not all points of a data set 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

\[
E[(y - B_{LS}x)'\Sigma_{LS}^{-1}(y - B_{LS}x)] = E[u'v] = \text{tr} E[\varepsilon'\varepsilon'] = \text{tr}(\Sigma_{LS}^{-1/2}E[\varepsilon'\varepsilon']\Sigma_{LS}^{-1/2}) = \text{tr} I_q = q,
\]

so \((B_{LS}, \Sigma_{LS})\) satisfies condition (A.1). Take any \( b \in \mathbb{R}^{p \times q} \) and any \( \Delta \) a positive definite symmetric matrix of size \( q \) such that (A.1) holds. There exists an orthogonal matrix \( P \) and \( \lambda_1 \geq \cdots \geq \lambda_q > 0 \) such that \( \Delta = \Sigma_{LS}^{1/2}PA^t\Sigma_{LS}^{1/2} \) where \( \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_q) \). Put \( v = P^t\Sigma_{LS}^{-1/2}(y - b'x) \). Then we obtain

\[
q = E[(y - b'x)^t\Delta^{-1}(y - b'x)] = E[v^t\Lambda^{-1}v] = \sum_{i=1}^q \lambda_i^{-1}E[v_i^2].
\]

(A.2)

On the other hand, since \( E[\varepsilon\varepsilon'] = 0 \), we have that

\[
E[v'v] = P^t\Sigma_{LS}^{1/2}E[\varepsilon + (B_{LS} - b)x](\varepsilon + (B_{LS} - b)x)'\Sigma_{LS}^{-1/2}P
\]

\[
= P^t(I_q + \Sigma_{LS}^{-1/2}(B_{LS} - b)'E[xx'](B_{LS} - b)\Sigma_{LS}^{-1/2})P
\]

\[
= I_q + ((B_{LS} - b)\Sigma_{LS}^{-1/2})^tE[xx']((B_{LS} - b)\Sigma_{LS}^{-1/2}).
\]

(A.3)

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

\[
q = \sum_{i=1}^q \lambda_i^{-1} + \sum_{i=1}^q \lambda_i^{-1}((B_{LS} - b)\Sigma_{LS}^{-1/2})^tE[xx']((B_{LS} - b)\Sigma_{LS}^{-1/2})i \geq \sum_{i=1}^q \lambda_i^{-1},
\]

(A.4)

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

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

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

Moreover, equality in (A.4) only occurs if all \(((B_{LS} - b)\Sigma_{LS}^{-1/2} P)\) = 0, thus if \( b = B_{LS} \). In order to have \( \det \Sigma_{LS} = \det \Delta \), also (A.5) needs to become an equality, which can only occur if all \( \lambda_i \) are equal to one, implying \( \Delta = \Sigma_{LS} \). Hereby, we have also proved the uniqueness part.

**Proof of Proposition 1.** Take \( \hat{H} \in \arg\min_{H} \det \hat{\Sigma}_{LS}(H) \). We first prove that \( \hat{B}_{LS}(\hat{H}) \) minimizes \( \det \text{MCD}_q(B) \). Take \( B \in \mathbb{R}^{p \times q} \) arbitrarily, then by definition of the MCD there exists a \( H \in \mathcal{H} \) such that \( \text{MCD}_q(B) = \text{Cov}_0(H, B). \) Using properties of traces, it follows that

\[
\frac{1}{h} \sum_{j \in H} r_j(B)(\text{Cov}_0(H, B))^{-1} r_j(B)' = q. \tag{A.6}
\]

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

\[
\det \text{MCD}_q(B) = \det \text{Cov}_0(H, B) \geq \det \hat{\Sigma}_{LS}(H) \geq \det \hat{\Sigma}_{LS}(\hat{H}) = \det \text{Cov}_0(\hat{H}, \hat{B}_{LS}(\hat{H})) \geq \det \text{MCD}_q(\hat{B}_{LS}(\hat{H})),
\]

where we applied the definition of \( \hat{H} \) and \( \text{MCD}_q \). We conclude that \( \hat{B}_{LS}(\hat{H}) \in \arg\min_{B} \det \text{MCD}_q(B) \).

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

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

Hence, we have that \( \hat{H} \in \arg\min_{H} \det \Sigma_{LS}(H) \) which ends the proof.

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

\[
\frac{1}{h} \sum_{j \in H} d_j^2(\hat{B}_{LS}(H), \hat{\Sigma}_{LS}(H)) = \frac{1}{h} \text{tr} \sum_{j \in H} \hat{\Sigma}_{LS}(H)^{-1} r_j(\hat{B}_{LS}(H)) r_j(\hat{B}_{LS}(H))' = \text{tr} \hat{\Sigma}_{LS}(H)^{-1} \hat{\Sigma}_{LS}(H) = q. \tag{A.7}
\]

We also have that

\[
\sum_{j \in H} d_j^2(\hat{B}_{LS}(H), \hat{\Sigma}_{LS}(H)) = (\det \hat{\Sigma}_{LS}(H))^{-1/q} \sum_{j \in H} d_j^2(\hat{B}_{LS}(H), \hat{\Sigma}_{LS}(H)). \tag{A.8}
\]

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

\[
\sum_{j \in H} d_j^2(\hat{B}_{LS}(H), \hat{\Sigma}_{LS}(H)) = h q \det \hat{\Sigma}_{LS}(H)^{1/q}. \tag{A.9}
\]
We first prove that for any \( \hat{H} \in \text{argmin} \det \hat{\Sigma}_{LS}(H) \) we have that \( \hat{B}_{LS}(\hat{H}) \in \{ \tilde{B} | (\tilde{B}, \tilde{\Sigma}) \in \text{argmin} \sum_{B, \Sigma:|\Sigma|=1} d_{j,n}^2(B, \Sigma) \} \). Take \( \hat{H} \in \text{argmin} \det \hat{\Sigma}_{LS}(H) \) and denote

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

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

\[
\sum_{j=1}^{h} d_{j,n}^2(\hat{B}_{LS}(\hat{H}), \hat{\Sigma}_{LS}(\hat{H})) = \sum_{j \in H'} d_{j}^2(\hat{B}_{LS}(\hat{H}), \hat{\Sigma}_{LS}(\hat{H})) < \sum_{j \in H} d_{j}^2(\hat{B}_{LS}(\hat{H}), \hat{\Sigma}_{LS}(\hat{H})).
\]

Using (A.8) and (A.9), this yields \( \frac{1}{h} \sum_{j \in H'} d_{j}^2(\hat{B}_{LS}(\hat{H}), \hat{\Sigma}_{LS}(\hat{H})) < q \). Therefore, there exists a constant \( 0 < c < 1 \) such that \( \frac{1}{h} \sum_{j \in H'} d_{j}^2(\hat{B}_{LS}(\hat{H}), \hat{\Sigma}_{LS}(\hat{H})) < q \). It then follows from Lemma 2 that \( \det \hat{\Sigma}_{LS}(H') < \det c \hat{\Sigma}_{LS}(\hat{H}) \) which is a contradiction, so we conclude that

\[
\sum_{j=1}^{h} d_{j,n}^2(\hat{B}_{LS}(\hat{H}), \hat{\Sigma}_{LS}(\hat{H})) = \sum_{j \in H} d_{j}^2(\hat{B}_{LS}(\hat{H}), \hat{\Sigma}_{LS}(\hat{H})).
\]  

(A.10)

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

\[
\sum_{j=1}^{h} d_{j,n}^2(B, \Sigma) < \sum_{j=1}^{h} d_{j,n}^2(\hat{B}_{LS}(\hat{H}), \hat{\Sigma}_{LS}(\hat{H})).
\]  

(A.11)

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

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

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

\[
\sum_{j=1}^{h} d_{j,n}^2(B, \Sigma) > \sum_{j \in H_1} d_{j}^2(\hat{B}_{LS}(H_1), \tilde{\Sigma}_{LS}(H_1)).
\]  

(A.12)

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

\[
\sum_{j \in H_1} d_{j}^2(\hat{B}_{LS}(H_1), \tilde{\Sigma}_{LS}(H_1)) < \sum_{j \in H} d_{j}^2(\hat{B}_{LS}(\hat{H}), \hat{\Sigma}_{LS}(\hat{H})).
\]  

(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 \(\hat{H} \in \arg\min_{\hat{H}} \det \hat{\Sigma}_{LS}(\hat{H})\).

Therefore, we conclude that

\[
\sum_{j=1}^{h} d_{j,n}^2(\hat{B}_{LS}(\hat{H}), \hat{\Sigma}_{LS}(\hat{H})) \leq \sum_{j=1}^{h} d_{j,n}^2(B, \Sigma)
\]

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

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

\[
\sum_{j=1}^{h} d_{j,n}^2(\hat{B}, \hat{\Sigma}) = \sum_{j \in \bar{H}} d_j^2(\hat{B}, \hat{\Sigma}) \leq \sum_{j \in \bar{H}} d_j^2(\hat{B}_{LS}(\hat{H}), \hat{\Sigma}_{LS}(\hat{H})).
\]

(A.14)

Using (A.9) it follows that \(\frac{1}{h} \sum_{j \in \bar{H}} d_j^2(\hat{B}, \det \hat{\Sigma}_{LS}(\hat{H})^{1/q} \hat{\Sigma}) \leq q\). Hence, there exists a constant \(0 < c \leq 1\) such that \(\frac{1}{h} \sum_{j \in \bar{H}} d_j^2(\hat{B}, c \det \hat{\Sigma}_{LS}(\hat{H})^{1/q} \hat{\Sigma}) = q\). From Lemma 2 we then obtain that \(\det \hat{\Sigma}_{LS}(\bar{H}) \leq \det (c \det \hat{\Sigma}_{LS}(\hat{H})^{1/q} \hat{\Sigma}) = c^q \det \hat{\Sigma}_{LS}(\hat{H})\) which is a contradiction unless if \(c = 1\) and by Lemma 2 (uniqueness) we then have that \(\hat{B} = \hat{B}_{LS}(\hat{H})\) and \(\hat{\Sigma} = \hat{\Sigma}_{LS}(\hat{H})\). For any \(H \in \mathcal{H}\) we now have that

\[
\sum_{j=1}^{h} d_{j,n}^2(\hat{B}, \hat{\Sigma}) = \sum_{j \in \bar{H}} d_j^2(\hat{B}_{LS}(\hat{H}), \hat{\Sigma}_{LS}(\hat{H})) \leq \sum_{j \in \bar{H}} d_j^2(\hat{B}_{LS}(H), \hat{\Sigma}_{LS}(H)).
\]

By using (A.9) the inequality can be rewritten as \(hq \det \hat{\Sigma}_{LS}(\bar{H})^{1/q} \leq hq \det \hat{\Sigma}_{LS}(H)^{1/q}\) which yields \(\det \hat{\Sigma}_{LS}(\bar{H}) \leq \det \hat{\Sigma}_{LS}(H)\) for all \(H \in \mathcal{H}\). Therefore, we conclude that \(\hat{H} \in \arg\min_{\hat{H}} \det \hat{\Sigma}_{LS}(\hat{H})\) which ends the proof. \(\square\)

**Proof of Theorem 1.** We first prove that \(\epsilon_n^2(\hat{B}_{MLTS}, Z_n) \geq \min(n - h + 1, h - k(Z_n))/n\). We will show that there exists a value \(\tilde{M}\), 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 \(\|\hat{B}_{MLTS}(Z'_n)\| \leq \tilde{M}\). 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 \(\varepsilon_1, \varepsilon_2 > 0\) such that \(\varepsilon_1 \|A\| \leq \|A\|_2 \leq \varepsilon_2 \|A\|\) 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 $B \in \mathbb{R}^{p \times q}$ such that $y_j - B' x_j$ for all $j \in J$ are lying on a $(q - 1)$-dimensional hyperplane. Indeed, otherwise there exists a $z \in \mathbb{R}^q$ such that for all $j \in J$ we have $z'(y_j - B' x_j) = z'y_j - \gamma' x_j = 0$ where $\gamma = Bz$. However, this contradicts the assumption $\# J = k(Z_n) + 1$. Since for all $B \in \mathbb{R}^{p+q}$ the $r_j := y_j - B' x_j$ are not lying on a $(q - 1)$-dimensional hyperplane, we have that

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

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

$$c = \min \left( \min(c_1(J), c_2(J)) \right) > 0,$$

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

$$M = \sup_{H \in \mathcal{H}} \|B_{LS}(H)\| < \infty$$

(A.16)

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 $V = (N_y + M N_x)^2 q$ and

$$\tilde{M} = \left( \left( \frac{h}{k(Z_n) + 1} c \right)^{1-q} + N_y \right) \frac{1}{4} c.$$  

(A.17)

Now take a data set $Z'_n$ obtained by replacing $m$ observations from $Z_n$ and suppose $\|\hat{B}_{MLTS}(Z'_n)\| > M$. First of all, there exists a subset $H_1 \in \mathcal{H}$ containing indices only corresponding to data points of the original data set $Z_n$. Using Lemma 5.1 in [16, p. 244] and properties of norms it follows that

$$\det(\hat{\Sigma}_{LS}(H_1)) \leq \lambda_{\max}(\text{cov}(r_j(\hat{B}_{LS}(H_1)); j \in H_1))^q$$

$$\leq \left( \frac{1}{h} \sum_{j \in H_1} \lambda_{\max}(r_j(\hat{B}_{LS}(H_1))r_j(\hat{B}_{LS}(H_1))^t) \right)^q$$

$$= \left( \frac{1}{h} \sum_{j \in H_1} \|r_j(\hat{B}_{LS}(H_1))\|^2 \right)^q$$

$$\leq \left( \frac{1}{h} \sum_{j \in H_1} (\|y_j\| + \|\hat{B}_{LS}(H_1)'x_j\|^2) \right)^q$$

$$\leq (N_y + M N_x)^{2q}$$

$$= V,$$  

(A.18)

where $\lambda_{\max}$ denotes the largest eigenvalue of a matrix. Now let $H_2$ be the optimal subset corresponding to $\hat{B}_{MLTS}(Z'_n)$ such that $\hat{B}_{MLTS}(Z'_n) = \hat{B}_{LS}(H_2) := B_2$. Since $h - m \geq k(Z_n) + 1$ the set $H_2$ contains a subset $\tilde{J}$ of size $k(Z_n) + 1$ corresponding to original observations of $Z_n$. Using
again Lemma 5.1 in [16] we obtain
\[
\lambda_{\min}(\hat{\Sigma}_{LS}(H_2)) = \lambda_{\min}(\text{cov}((y_j - B'_2 x_j; j \in H_2))) \\
\geq \frac{k(Z_n) + 1}{h} \lambda_{\min}(\text{Cov}_0((y_j - B'_2 x_j; j \in \tilde{J}))) \\
\geq \frac{k(Z_n) + 1}{c_2(\tilde{J})} \\
\geq \frac{k(Z_n) + 1}{c}.
\]
(A.19)

On the other hand,
\[
\lambda_{\max}(\hat{\Sigma}_{LS}(H_2)) = \sup_{\|u\|=1} \frac{1}{h} \sum_{j \in H_2} u'(y_j - B'_2 x_j)(y_j - B'_2 x_j)'u.
\]
(A.20)

By definition of \(c_1(\tilde{J})\) there exists at least one index \(j_0 \in \tilde{J} \subset H_2\) such that
\[
\|B'_2 x_{j_0}\|^2 = \sum_{j=1}^q |B_{2j} x_{j_0}|^2 \geq \sum_{j=1}^q \|B_{2j}\|^2 c_1(\tilde{J})^2 = (\|B_{2}\| c_1(\tilde{J}))^2 \geq (x_1 \|B_{2}\| c_1(\tilde{J}))^2
\]
which yields \(\|B'_2 x_{j_0}\| > x_1 \tilde{M} c\). Since by definition \(x_1 \tilde{M} c \geq N_y\) we obtain \(\|y_{j_0} - B'_2 x_{j_0}\| \geq \|y_{j_0}\| - \|B'_2 x_{j_0}\| > x_1 \tilde{M} c - N_y\). By taking \(u = \frac{y_{j_0} - B'_2 x_{j_0}}{\|y_{j_0} - B'_2 x_{j_0}\|}\) it follows from (A.20) that
\[
\lambda_{\max}(\hat{\Sigma}_{LS}(H_2)) \geq \|y_{j_0} - B'_2 x_{j_0}\|^2 / h > (x_1 \tilde{M} c - N_y)^2 / h.
\]
(A.21)

Combining (A.21) and (A.19) yields
\[
\det(\hat{\Sigma}_{LS}(H_2)) > \frac{1}{h} \left( x_1 \tilde{M} c - N_y \right)^2 \left( \frac{k(Z_n) + 1}{c} \right)^{q-1} = V
\]
by definition of \(\tilde{M}\). Together with (A.18) this implies \(\det(\hat{\Sigma}_{LS}(H_2)) > \det(\hat{\Sigma}_{LS}(H_1))\) which contradicts the definition of \(\hat{B}_{MLTS}(Z_n)\), so we conclude that \(\|\hat{B}_{MLTS}(Z_n)\| \leq \tilde{M}\).

We now prove that also \(\varepsilon_n^*(\hat{B}_{MLTS}, Z_n) \leq \min(n - h + 1, h - k(Z_n))/n\). 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 \(\hat{B}_{MLTS}(Z_n)\) explodes.

Now we show that \(\varepsilon_n^*(\hat{B}_{MLTS}, Z_n) \leq (h - k(Z_n))/n\). 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 \(\beta' y_j - \gamma' x_j = 0\) for all \(j \in \tilde{J}\).

If \(\beta \neq 0\) then there exists a \(B \in \mathbb{R}^{p+q}\) such that \(B\beta = \gamma\) which implies \(\beta' (y_j - B' x_j) = 0\) for \(j \in \tilde{J}\). Therefore, for \(j \in \tilde{J}\) we have that \(y_j - B' x_j \in S\) where \(S\) is a \((q-1)\)-dimensional subspace of \(\mathbb{R}^q\). Now take a \(D \in \mathbb{R}^{p+q}\) with \(\|D\| = 1\) such that \(\{D' x; x \in \mathbb{R}^p\} \subset S\). Now replace \(m = h - k(Z_n)\) observations of \(Z_n\), not lying on \(S\), by \((x_0, (B + \lambda D)' x_0)\) for some arbitrarily chosen \(x_0 \in \mathbb{R}^p\) and \(\lambda \in \mathbb{R}\). Denote \(J_0\) the set of indices corresponding to the outliers. It follows that for the \(m\) outliers \(r_j(B + \lambda D) = 0\) and for the \(k(Z_n)\) points on \(S\) we have that \(r_j(B + \lambda D) = y_j - B' x_j - \lambda D' x_j \in S\). Therefore, \(\{r_j(B + \lambda D); j \in \tilde{J} \cup J_0\}\) belongs to the subspace \(S\), giving a zero determinant for the matrix \(\text{cov}_0(r_j(B + \lambda D); j \in \tilde{J} \cup J_0)\). Therefore, using Proposition 1 it follows that \(\hat{B}_{MLTS}(Z_n) = B + \lambda D\) which tends to infinity when \(\lambda \to \infty\).
If $\beta = 0$ then we have that $\gamma^i x_j = 0$ for all $j \in \tilde{J}$. Now replace $m = h - k(Z_n)$ other observations of $Z_n$ by observations on the hyperplane $\gamma^i x = 0$. Denote $H_2$ the set of indices corresponding with observations of $Z_n$ such that $\gamma^i x = 0$. Since all these observations belong to a hyperplane of $\mathbb{R}^{p+q}$ we have that $\det\text{cov}((y_j - \hat{B}_{\text{LS}}(H_2)^i x_j; j \in H_2)) = 0$. But since $\gamma^i x = 0$ is a vertical hyperplane we have $\|\hat{B}_{\text{LS}}(H_2)\| = \infty$ and it follows that $\|\hat{B}_{\text{MLTS}}(Z_n)\| = \infty$. □

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

**Proof of Corollary 2.** Since for $q = 1$ we have $\det(\hat{\Sigma}_{\text{LS}}(H_2)) = \lambda_{\text{max}}(\hat{\Sigma}_{\text{LS}}(H_2))$, we do not need to establish the lower bound (A.19) and thus we do not need $c_2(\tilde{J}) > 0$. To obtain $c_1(\tilde{J}) > 0$ it suffices to consider data sets 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 Corollary 3.** Denote $Z_n = \{(1, y_i); 1 \leq i \leq n\}$ then clearly $\hat{\mu}_{\text{MCD}}(Y_n) = \hat{B}_{\text{MLTS}}(Z_n)$ and $\hat{\Sigma}_{\text{MCD}}(Y_n) = \hat{\Sigma}_{\text{MLTS}}(Z_n)$. 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. □

**Proof of Theorem 2.** Using properties of traces we obtain

$$\frac{1}{h} \sum_{j \in H_2} d^2_j(\hat{B}_2, \hat{\Sigma}_2) = \frac{1}{h} \text{tr} \sum_{j \in H_2} r_j(\hat{B}_2)^i \hat{\Sigma}_2^{-1} r_j(\hat{B}_2) = \text{tr} \hat{\Sigma}_2^{-1} \hat{\Sigma}_2 = \text{tr} I_q = q$$

(A.22)

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

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

(A.23)

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

$$\frac{1}{h} \sum_{j \in H_2} r_j(\hat{B}_1)^i (c \hat{\Sigma}_1)^{-1} r_j(\hat{B}_1) = \frac{1}{eh} \sum_{j \in H_2} d^2_j(\hat{B}_1, \hat{\Sigma}_1) = \frac{cq}{c} = q.$$  

(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 $\hat{B}_2 = \hat{B}_1$ 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 $\hat{B}_2 = \hat{B}_1$ and $\hat{\Sigma}_2 = \hat{\Sigma}_1$. □

**Proof of Lemma 1.** Clearly, we have that $E \in D_H(a)$. Note that

$$\frac{1}{1 - a} \int_\hat{\Delta} d^2(x, y) dH = \frac{1}{1 - a} \text{tr} \int_\hat{\Delta} d^2(x, y) dH = \text{tr}(\hat{\Sigma}_A(H)^{-1} \hat{\Sigma}_A(H)) = \text{tr} I_q = q.$$
On the other hand, we have that
\[
\int_E d^2(x, y) dH = \int_{E \cap \hat{A}} d^2(x, y) dH + \int_{E \setminus \hat{A}} d^2(x, y) dH
\leq \int_{E \cap \hat{A}} d^2(x, y) dH + D_H^2 \mathbb{P}_H(\hat{A} \setminus E)
= \int_{E \cap \hat{A}} d^2(x, y) dH + \int_{\hat{A} \setminus E} d^2(x, y) dH
= \int_{\hat{A}} d^2(x, y) dH.
\]
Therefore, there exists a \(0 < c \leq 1\) such that
\[
\frac{1}{1 - x} \int_C (y - (B_{\hat{A}}(H))^t x)^t (c \Sigma_{\hat{A}}(H))^{-1} (y - (B_{\hat{A}}(H))^t x) dH = q.
\] (A.25)

Since \(\hat{A}\) is an MCD solution, we have that \(\det(c \Sigma_{\hat{A}}(H)) \leq \det \Sigma_{\hat{A}}(H) \leq \det \Sigma_E(H)\) which in combination with (A.25) contradicts Lemma 2 unless if \(B_{\hat{A}}(H) = B_E(H)\) and \(c \Sigma_{\hat{A}}(H) = \Sigma_E(H)\). Then \(c\) should also be equal to 1. □

**Proof of Theorem 3.** First of all, due to equivariance, we may assume that \(B = 0\) and \(\Sigma = I_q\), so \(y = \varepsilon \sim F\). It now suffices to show that \(B_{MLTS}(H) = 0\). Then we will have that \(\Sigma_{MLTS}(H)\) is the MCD functional at the distribution of \(y - B_{MLTS}(H)^t x = y - \varepsilon\). Since the factor \(c_x\) makes the MCD Fisher-consistent at elliptical distributions (see [3,4]) it will follow that \(\Sigma_{MLTS}(H) = I_q\).

Lemma 1 shows that \(B_{MLTS}\) is the least-squares fit based solely on the cylinder \(C = \{(x, y) \in \mathbb{R}^{p+q}; (y - B_{MLTS}^t x)^t \Sigma_{MLTS}^{-1} (y - B_{MLTS}^t x) \leq D_H^2\}\). Therefore,
\[
\int_C x(y - B_{MLTS}^t x)^t dH(x, y) = 0.
\] (A.26)

Now suppose that \(B_{MLTS} \neq 0\). Let \(\lambda_1, \ldots, \lambda_q\) be the eigenvalues of \(\Sigma_{MLTS}\) and \(v_1, \ldots, v_q\) the corresponding eigenvectors. There will be at least one \(1 \leq j \leq q\) such that \(B_{MLTS} v_j \neq 0\). (Note that \(B_{MLTS}\) is not necessarily of full rank.) Fix this \(j\). From (A.26) it follows that we should have
\[
\int_E v_j^t (B_{LTS}^t x)(y - B_{MLTS}^t x)^t v_j dF(y) dG(x) = 0
\]
which can be rewritten as
\[
\int_{\mathbb{R}^p} v_j^t (B_{MLTS}^t x) I(x) dG(x) = 0
\] (A.27)
with
\[
I(x) = \int_C (y - B_{MLTS}^t x)^t v_j dF(y),
\]
where \(C_x = \{y \in \mathbb{R}^q | (x, y) \in C\}\). Fix \(x\) and set \(d = (d_1, \ldots, d_q) := B_{MLTS}^t x\). Since \(y\) is spherically symmetrically distributed, for computing \(I(x)\) we may assume w.l.o.g. that
\[ \Sigma_{\text{MLTS}} = \text{diag}(\lambda_1, \ldots, \lambda_q) \] 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

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

where \( c := D_2^2 > 0 \). Then we can rewrite \( I(x) \) as

\[
I(x) = \int_{d_1 - \sqrt{c \lambda_1}}^{d_1 + \sqrt{c \lambda_1}} \int_{C(y_1)} (y_1 - d_1)g(y_1^2 + \cdots + y_q^2) dy_1 \ldots dy_q \\
= \int_{-\sqrt{c \lambda_1}}^{\sqrt{c \lambda_1}} t \int_{C(d_1 + t)} g((d_1 + t)^2 + y_2^2 + \cdots + y_q^2) dy_2 \ldots dy_q dt.
\]

Since \( C(d_1 + t) = C(d_1 - t) \) it follows that

\[
I(x) = \int_{0}^{\sqrt{c \lambda_1}} t \int_{C(d_1 + t)} g \left((d_1 + t)^2 + y_2^2 + \cdots + y_q^2\right) \\
- g \left((d_1 - t)^2 + y_2^2 + \cdots + y_q^2\right) dy_2 \ldots dy_q dt.
\]

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 \( v_j^i(B_{\text{MLTS}}^i) > 0 \) implies \( I(x) < 0 \) and if \( v_j^i(B_{\text{MLTS}}^i) < 0 \), then \( I(x) > 0 \). Also, \( v_j^i(B_{\text{MLTS}}^i) = 0 \) implies \( I(x) = 0 \). However, due to condition (24), the latter event occurs with probability less than 1 \( - x \). Therefore, we obtain \( \int_{\mathbb{R}^q} v_j^i B_{\text{MLTS}}^i I(x) dG(x) < 0 \) which contradicts (A.27), so we conclude that \( B_{\text{MLTS}} = 0 \). \( \square \)

**Proof of Theorem 4.** Consider the contaminated distribution \( H_\varepsilon = (1 - \varepsilon) H_0 + \varepsilon A_{z_0} \) with \( z_0 = (x_0, y_0) \) and denote \( B_\varepsilon := B_{\text{MLTS}}(H_\varepsilon) \) and \( \Sigma_\varepsilon := \Sigma_{\text{MLTS}}(H_\varepsilon) \). Then (20) results in

\[
\hat{B}_\varepsilon = \left( \int_{\hat{A}_\varepsilon} xx' dH_\varepsilon(x, y) \right)^{-1} \int_{\hat{A}_\varepsilon} xy' dH_\varepsilon(x, y),
\]

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

\[
IF(z_0; B_{\text{MLTS}}, H_0) = \left( \int_{\hat{A}_\varepsilon} xx' dH_0(z) \right)^{-1} \frac{\partial}{\partial \varepsilon} \int_{\hat{A}_\varepsilon} xy' dH_\varepsilon(z)|_{\varepsilon=0}
\
+ \frac{\partial}{\partial \varepsilon} \left[ \left( \int_{\hat{A}_\varepsilon} xx' dH_\varepsilon(z) \right)^{-1} \right]|_{\varepsilon=0} \int_{\hat{A}_\varepsilon} xy' dH_0(z)
\]

Lemma 1 combined with Fisher-consistency yields that \( \hat{A} = \{(x, y) \in \mathbb{R}^{p+q}; y'y \leq q_x \} \) where \( q_x = (D_2^p)^{-1}(1 - x) \) with \( D_2^p(t) = P_F(\|y\|^2 \leq t) \). Hence \( \hat{A} = \mathbb{R}^p \times \{y \in \mathbb{R}^q; \|y\|^2 \leq q_x \} =: \mathbb{R}^p \times A \). This implies

\[
\int_{\hat{A}} xy' dH_0(z) = \int_{\mathbb{R}^p} x dG(x) \int_{\hat{A}} y' dF(y) = 0
\]
by symmetry of $F$ and

$$
\int_{A} xx' \, dH_0(z) = \int_{\mathbb{R}^p} xx' \, dG(x) \int_{A} \, dF(y) = EG[xx'] \, (1 - \alpha).
$$

Therefore, we obtain

$$
IF(z_0; B_{MLTS}, H_0) = \frac{E_G[xx']^{-1}}{1 - \alpha} \left( \frac{\partial}{\partial \varepsilon} \int_{\hat{A}_{\varepsilon}} xx' \, dH_0(z) \right)_{\varepsilon = 0}
$$

$$
= \frac{E_G[xx']^{-1}}{1 - \alpha} \left( (1 - \varepsilon) \int_{\hat{A}_{\varepsilon}} xx' \, dH_0(z) + \varepsilon x_0 y_0^t I(z_0 \in \hat{A}_{\varepsilon}) \right)_{\varepsilon = 0}
$$

$$
= \frac{E_G[xx']^{-1}}{1 - \alpha} \left( x_0 y_0^t I(\|y_0\|^2 \leq q_2) + \frac{\partial}{\partial \varepsilon} \int_{\hat{A}_{\varepsilon}} xx' \, dH_0(z) \right).
$$

Similarly to Proposition 1 in [4], it can be shown that Lemma 1 still holds for contaminated distributions $H_0$. Let us denote $d_2^2(x, y) = (y - B_2' x)^t \Sigma_2^{-1} (y - B_2' x)$, then it follows that $\hat{A}_{\varepsilon} = \{(x, y) \in \mathbb{R}^{p+q}; d_2^2(x, y) \leq q_2(\varepsilon)\}$ where $q_2(\varepsilon) = (D_2^2 H_0)^{-1}(1 - \varepsilon) \text{ with } D_2^2 H_0(\tau) = P_{H_0}(d_2^2(x, y) \leq \tau)$. For $x$ fixed we define the ellipsoid $E_{\varepsilon,x} := \{y \in \mathbb{R}^q; d_2^2(x, y) \leq q_2(\varepsilon)\}$. Then it follows that

$$
\int_{\hat{A}_{\varepsilon}} xx' \, dH_0(z) = \int_{\mathbb{R}^p} \int_{E_{\varepsilon,x}} xx' \, dF(y) \, dG(x)
$$

$$
= \int_{\mathbb{R}^p} x \left( \int_{E_{\varepsilon,x}} y \, g(y'y) \, dy \right)^t \, dG(x).
$$

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

$$
I(\varepsilon) := \int_{E_{\varepsilon,x}} y \, g(y'y) \, dy = \det(\Sigma_\varepsilon)^{1/2} \int_{\|v\|^2 \leq q_2(\varepsilon)} (\Sigma_\varepsilon^{1/2} v + B_\varepsilon' x)
$$

$$
\times g((\Sigma_\varepsilon^{1/2} v + B_\varepsilon' x)^t (\Sigma_\varepsilon^{1/2} v + B_\varepsilon' x)) \, dv.
$$

Rewriting this expression in polar coordinates $v = re(\theta)$ where $r \in [0, \sqrt{q_2(\varepsilon)}]$, $e(\theta) \in S^{q-1}$ and $\theta = (\theta_1, \ldots, \theta_{q-1}) \in \Theta = [0, \pi] \times \cdots \times [0, \pi]$, yields

$$
I(\varepsilon) = \det(\Sigma_\varepsilon)^{1/2} \int_{0}^{\sqrt{q_2(\varepsilon)}} \int_{\Theta} J(\theta, r)(r^{1/2} e(\theta) + B_\varepsilon' x)
$$

$$
\times g((r^{1/2} e(\theta) + B_\varepsilon' x)^t (r^{1/2} e(\theta) + B_\varepsilon' x)) \, dr \, d\theta,
$$

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

$$
\frac{\partial}{\partial \varepsilon} I(\varepsilon)_{\varepsilon = 0} = \int_{\|v\|^2 \leq q_2} \frac{\partial}{\partial \varepsilon} \left( (\Sigma_\varepsilon^{1/2} v + B_\varepsilon' x) g((\Sigma_\varepsilon^{1/2} v + B_\varepsilon' x)^t (\Sigma_\varepsilon^{1/2} v + B_\varepsilon' x)) \right)_{\varepsilon = 0} \, dv.
$$

The derivative on the right-hand side becomes

$$
\frac{\partial}{\partial \varepsilon} \left((\Sigma_\varepsilon^{1/2} v + B_\varepsilon' x) g((\Sigma_\varepsilon^{1/2} v + B_\varepsilon' x)^t (\Sigma_\varepsilon^{1/2} v + B_\varepsilon' x))\right)_{\varepsilon = 0}
$$

$$
= IF(z_0; \Sigma_{MLTS}^{1/2}, H_0) v + IF(z_0; B_{MLTS}, H_0)^t x g(v'v)
$$

$$
+ 2 v g'(v'v) (v' IF(z_0; \Sigma_{MLTS}^{1/2}, H_0) v + v' IF(z_0; B_{MLTS}, H_0)^t x).
$$
Since $\int_{\|v\|^2 \leq q_2} v g(v^t v) dv$ and $\int_{\|v\|^2 \leq q_2} v g'(v^t v)v^t IF(z_0; \Sigma_{\text{MLTS}}^{1/2}, H_0) v dv$ are zero due to symmetry of $F$, the terms in (A.31) including $IF(z_0; \Sigma_{\text{MLTS}}^{1/2}, H_0)$ give a zero contribution to the integral in (A.30). It follows that
\[
\frac{\partial}{\partial \varepsilon} \text{I}(\varepsilon)|_{\varepsilon = 0} = (1 - \zeta)IF(z_0; B_{\text{MLTS}}, H_0)'x + 2 \int_{\|v\|^2 \leq q_2} g'(v^t v)v^t dv IF(z_0; B_{\text{MLTS}}, H_0)'x
\]
\[
= [(1 - \zeta) + 2c_2]IF(z_0; B_{\text{MLTS}}, H_0)'x,
\]
where $c_2 = \int_{\|v\|^2 \leq q_2} g'(v^t v)v^t dv$ can be rewritten in the form given in Theorem 4 by using polar coordinates. From (A.29) we now obtain that
\[
\frac{\partial}{\partial \varepsilon} \int_{A_\varepsilon} xy^t dH_0(z)|_{\varepsilon = 0} = [(1 - \zeta) + 2c_2]E_G[xx^t]IF(z_0; B_{\text{MLTS}}, H_0).
\]
(A.32)
Substituting (A.32) in (A.28) yields
\[
(1 - \zeta)IF(z_0; B_{\text{MLTS}}, H_0) = E_G[xx^t]^{-1}xy^t I(\|y\|^2 \leq q_2)
\]
\[
+ [(1 - \zeta) + 2c_2]IF(z_0; B)_{\text{MLTS}}, H_0)
\]
which results in
\[
IF(z_0; B_{\text{MLTS}}, H_0) = E_G[xx^t]^{-1}xy^t I(\|y\|^2 \leq q_2).
\]

References