Noname manuscript No.
(will be inserted by the editor)

Optimization techniques for multivariate least trimmed absolute deviation estimation

G. Zioutas · C.Chatzinakos · T.D. Nguyen · L. Pitsoulis*

the date of receipt and acceptance should be inserted later

Abstract Given a dataset an outlier can be defined as an observation that does not follow the statistical properties of the majority of the data. Computation of the location estimate is of fundamental importance in data analysis, and it is well known in statistics that classical methods, such as taking the sample average, can be greatly affected by the presence of outliers in the data. Using the median instead of the mean can partially resolve this issue but not completely. For the univariate case, a robust version of the median is the Least Trimmed Absolute Deviation (LTAD) robust estimator introduced in [18], which has desirable asymptotic properties such as robustness, consistently, high breakdown and normality. There are different generalizations of the LTAD for multivariate data, depending on the choice of norm. In [5] we present such a generalization using the Euclidean norm and propose a solution technique for the resulting combinatorial optimization problem, based on a necessary condition, that results in a highly convergent local search algorithm. In this subsequent work, we use the L^1 norm to generalize the LTAD to higher dimensions, and show that the resulting mixed integer programming problem has an integral relaxation, after applying an appropriate data transformation. Moreover, we utilize the structure of the problem to show that the resulting LP's can be solved efficiently using a subgradient optimization approach. The robust statistical properties of the proposed estimator are verified by extensive computational results.

Keywords: robust location estimation; least trimmed absolute deviation; outlier detection; linear programming; mixed integer programming.

1 Introduction

The sample average and standard deviation are the classical estimators of the location and the scale parameters of a statistical distribution. It is well-known that these classical estimators, although being optimal under normality assumptions, are extremely sensitive to the presence of outliers in the data;

* work of this author was conducted at National Research University Higher School of Economics and supported by RSF grant 14-41-00039

C. Chatzinakos · L. Pitsoulis · G. Zioutas
Department of Electrical and Computer Engineering
Aristotle University of Thessaloniki
Greece

T.D. Nguyen
Southampton Management School
University of Southampton, U.K

a small proportion of outliers in the data can have a large distorting effect on the sample mean and covariance. Robust statistics is concerned with the development of methods for computing estimators that are justifiably resistant to the presence of outliers in the data. The focus of this work is to estimate the unknown location parameter $\mathbf{m}$ of a family of distributions $F_{\mathbf{m}}$ given some data contaminated with an unknown number of outliers.

Detecting outliers and unusual data structures is one of the main problems in statistical data analysis since this occurs in many different application domains. One such application is in large scale complex networks, where the graph data may arrive in streams [1, 2]. Given that the degrees in social networks typically follow a power law distribution, it is of importance to identify outlier nodes which do not follow the degree distribution of the majority of the nodes. These outliers may affect the computation of several graph characteristics such as community detection, clustering coefficient etc.

Projection pursuit is one of the typical approaches for outlier detection. The idea is to repeatedly project the multivariate data into the univariate space since univariate outlier detection is much simpler to handle by applying order statistics and visualization. Such methods are usually computationally intensive, but they are particularly useful for high-dimensional data with small sample size. One such technique is the principal components analysis in Filzmoser et al. [6]. Outlier detection can also be done based on estimations of the covariance matrix. The idea is to use estimated covariance structure in order to find a distance, usually the well-known Mahalanobis distance, from each observation to the center of the data cloud. One such method is the Minimum Covariance Determinant (MCD) introduced in Rousseeuw [15] and in Rousseeuw and Driessen [16]. Desirable properties for an estimator include high breakdown values, high efficiency, and fast computation. One famous robust location estimator is the multivariate Least Euclidean Distance (LED) as studied in Hettmansperger [7]. Other methods for robust location estimation include the transformation median (Chakraborty et al. [4]) and the Oja Multivariate half samples Median (HOMM) Oja [12]. The corresponding univariate cases of the half samples (HOMM) and MCD are the Least Trimmed Absolute Deviation (LTAD) and Least Trimmed Squared (LTS) estimators, respectively.

The LTAD robust estimator for univariate data was introduced in [18], where it was shown that to have desirable asymptotic properties such as robustness, consistently, high breakdown and normality. Moreover, in [18], the author also presents an algorithm to efficiently compute the LTAD in $O(n \log n)$ time. These methods however, do not generalize to higher dimensions. In [5] the LTAD is generalized to handle multivariate data using the Euclidean norm, and the resulting combinatorial optimization problem is solved by an approximate fixed-point like iterative procedure. Computational experiments in [5] on both real and artificial data indicate that the proposed method efficiently identifies both location and scatter outliers in varying dimensions and high degree of contamination. In this work we extend the results in [5], and present a different generalization of LTAD which is based on the L^1 norm. It is shown that the linear programming relaxation of the resulting mixed integer program is integral, after applying an appropriate equidistance data transformation. This implies that the LTAD can be computed as a series of linear programs, each of which can be solved efficiently using a subgradient optimization approach.

The rest of this paper is structured as follows. In Section 2 we present the generalized LTAD for multivariate data using the L^1 norm while its mixed integer programming formulation is given in Section 3. The integrality of the relaxation is presented in Section 4, along with the procedure to perform data transformation. The subgradient optimization approach for solving the resulting linear programs is presented in Section 5. Finally, in Section 6 we perform computational experiments in real and simulated data to compare the performance of our method with the one in [5].

2 Least trimmed absolute deviation estimator

Given a sample of n univariate observations $X_n = \{x_1, x_2, \dots, x_n\}$ where $x_i \in \mathbb{R}, i = 1, \dots, n$, we can state the well known location parameter *median* as follows:

$$m(X_n) = \arg\min_m \sum_{i=1}^{n} |x_i - m| \tag{1}$$

Let the $(\cdot)$ operator denote the order of the data points, i.e.,

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)},$$

then a trivial solution to (1) is

$$m(X_n) = \begin{cases} x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n+1}{2})}\right) & \text{if } n \text{ is even.} \end{cases}$$

The *least median of squares* [17] is the midpoint of the subset that contains half of the observations on each side. Three statistically desirable properties of an estimator are *equivariance*, *monotonicity* and *50% breakdown point*. Equivariance implies that if the data points are scaled and shifted then the value of the estimator will change accordingly, while monotonicity implies that the estimator cannot decrease in value if an observation increases. An estimator has 50% breakdown point if its value will be bounded for any arbitrary change of less than half of the observations. Basset [3] has proven that the median is the only estimator that satisfies all the three aforementioned properties.

If we make the assumption that $n - h$ of observations are outliers, where $h > \lceil n/2 \rceil$, we can define a robust version of the median which will be called *least trimmed absolute deviations* (LTAD) estimator, defined by the following problem:

$$\begin{aligned} m(X_n, h) = \arg\min_{m,T} \ & \sum_{x \in T} |x - m| \\ \text{s.t.} \ & |T| = h \\ & T \subseteq X_n \end{aligned} \tag{2}$$

which implies that we have to find that subset T of h observations out of n which have the least median value. In order to satisfy the high breakdown property, the value of h is set to $\lceil n/2 \rceil$. Solving (2) by complete enumeration would require the computation of the median for all possible $\binom{n}{h}$ subsets $T \subseteq X_n$ and choosing the one with the minimum value, which is computationally infeasible even for moderate values of n. The LTAD was introduced by Tableman [18] for fixed $h = \lceil n/2 \rceil$, where in addition to showing favorable theoretical properties the author also provided a simple procedure for its computation based on the observation that the solution to (2) will be the median of one of the following $(n - h)$ contiguous subsets

$$\{x_{(1)}, \dots, x_{(h)}\}, \{x_{(2)}, \dots, x_{(h+1)}\}, \dots, \{x_{(n-h)}, \dots, x_{(n)}\}.$$

Therefore, it suffices to compute the $(n - h)$ median values for the above subsets and choose the one which minizes the sum in (2). This process will require $O(n \log n)$ time to order the data points according to increasing value.

Consider now the multidimensional version of the LTAD defined in (2), where we have p-variate observations $X_n = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ with $\mathbf{x}_i \in \mathbb{R}^p, i = 1, \dots, n$. Moreover, without loss of generality we can assume that the observations are rescalled. The multivariate LTAD is defined as

$$\begin{aligned} \text{LTAD} : \mathbf{m}(X_n, h) = \arg\min_{\mathbf{m},T} \ & \sum_{\mathbf{x} \in T} ||\mathbf{x} - \mathbf{m}||_1 \\ \text{s.t.} \ & |T| = h \\ & T \subseteq X_n \end{aligned} \tag{3}$$

where $\|\cdot\|_1$ stands for the one norm, i.e., $\|\mathbf{x}-\mathbf{m}\|_1 = \sum_{i=1}^{p} |x_i - m_i|$. For the ease of exposition in the rest of the paper, we refer to both the univariate and the multivariate LTAD as the *LTAD problem*.

The LTAD problem can be approximated by an iterative algorithm similar to the procedure described in [5] for solving the related *least trimmed Euclidean distances* (LTED) estimator, which is defined as the LTAD in (3) with the only exception that the euclidean norm is used instead of the one norm. However, although this algorithm is very fast, it almost always converges to a local optimum of unknown quality.

In this paper we present a different solution method for the LTAD, by approximating its natural mixed integer nonlinear programming formulation with a mixed integer linear program whose linear programming relaxation is integral. We also develop specialized efficient solution methods for the resulting linear program, since the iterative nature of the proposed method requires multiple calls for solving them.

3 Mixed integer programming formulation

The LTAD estimate in (3) can be equivalently stated as the following mixed integer nonlinear programming problem

$$\begin{aligned} \text{MINLP-LTAD}: \quad \min_{\mathbf{w},\mathbf{m}} \quad & \sum_{i=1}^{n} w_i \|\mathbf{x}_i - \mathbf{m}\|_1 \\ \text{s.t.} \quad & \sum_{i=1}^{n} w_i = h \\ & w_i \in \{0,1\}, i = 1,\ldots,n. \end{aligned} \tag{4}$$

where the zero-one weights $\mathbf{w} = (w_1,\ldots,w_n)$ indicate whether observation i is an outlier ($w_i = 0$) or a good observation ($w_i = 1$). For any feasible tuple $(\mathbf{w},\mathbf{m})$ to (4), let $\mathbf{x}_{(i)}$ denote the vector $\mathbf{x} \in X_n$ with the i-th smallest $\|\mathbf{x}-\mathbf{m}\|_1$ value, and $w_{(i)}$ its corresponding weight. We can now write (4) as follows

$$\begin{aligned} \sum_{i=1}^{n} w_i \|\mathbf{x}_i - \mathbf{m}\|_1 &= \sum_{i=1}^{h} \left\|\mathbf{x}_{(i)} - \mathbf{m}\right\|_1 \\ &= \sum_{i=1}^{h} \left\|w_{(i)}\mathbf{x}_{(i)} - \mathbf{m}\right\|_1 \\ &= \sum_{i=1}^{n} \|w_i\mathbf{x}_i - \mathbf{m}\|_1 - (n-h)\|\mathbf{m}\|_1 . \end{aligned}$$

since $w(i) = 1$ for all $i = 1,\ldots,h$. Observe that as $\mathbf{m}$ approaches zero, then $\sum_{i=1}^{n} \|w_i\mathbf{x}_i - \mathbf{m}\|_1$ approaches $\sum_{i=1}^{n} w_i \|\mathbf{x}_i - \mathbf{m}\|_1$, thus, for small values of $\mathbf{m}$ problem (4) can be approximated by the following

$$\begin{aligned} \min_{\mathbf{w},\mathbf{m}} \quad & \sum_{i=1}^{n} \|w_i\mathbf{x}_i - \mathbf{m}\|_1 \\ \text{s.t.} \quad & \sum_{i=1}^{n} w_i = h \\ & w_i \in \{0,1\}, i = 1,\ldots,n. \end{aligned}$$

which is equivalent to the following mixed integer linear program

$$\begin{aligned}
\text{MILP-LTAD}: \min_{\mathbf{w},\mathbf{m}} \ & \sum_{i=1}^{n}\sum_{j=1}^{p} d_{ij} \qquad (5)\\
\text{s.t.} \ & \sum_{i=1}^{n} w_i = h\\
& w_i x_{ij} - m_j - d_{ij} \le 0, \quad i = 1,\dots,n, j = 1,\dots,p\\
& -w_i x_{ij} + m_j - d_{ij} \le 0, \quad i = 1,\dots,n, j = 1,\dots,p\\
& \mathbf{w} \in \{0,1\}^n
\end{aligned}$$

where $\mathbf{D}$ is an $n \times p$ matrix of auxiliary variables $d_{ij} = |w_i x_{ij} - m_j|$, and $X = (x_{ij})$ is the $n \times p$ observations matrix whose rows are $\mathbf{x}_i^T$ for $i = 1,\dots,n$.

There are two issues with the approximation of (4) with (5). First of all, we need to ensure that MILP-LTAD is a good approximation of the MINLP-LTAD. Secondly, we need to be able to solve (5) efficiently. We will resolve the first issue by iteratively transforming the data such that the optimal $\mathbf{m}$ approaches zero. For the second issue we will show that the resulting mixed integer linear programming problem is equivalent to a linear programming problem under certain assumptions.

4 Data transformation

Let us denote with LP-LTAD the linear programming relaxation of (5) where $\mathbf{w} \in [0,1]^n$, Consider the linear programming relaxation of (5)

$$\begin{aligned}
\text{LP-LTAD}: \min_{\mathbf{w},\mathbf{m}} \ & \sum_{i=1}^{n}\sum_{j=1}^{p} d_{ij} \qquad (6)\\
\text{s.t.} \ & \sum_{i=1}^{n} w_i = h\\
& w_i x_{ij} - m_j - d_{ij} \le 0, \quad i = 1,\dots,n, j = 1,\dots,p\\
& -w_i x_{ij} + m_j - d_{ij} \le 0, \quad i = 1,\dots,n, j = 1,\dots,p\\
& \mathbf{w} \in [0,1]^n
\end{aligned}$$

Let $(\mathbf{w}_{LP}^*, \mathbf{m}_{LP}^*)$ be the optimal solution of LP-LTAD. If $\mathbf{w}_{LP}^*$ is integer, then this LP solution is also an optimal solution of (5).

We show next, that if in the linear programming optimal solution $\mathbf{m}_{LP}^*$ is equal to zero, then $(\mathbf{w}_{LP}^*, \mathbf{m}_{LP}^*)$ is optimal for the MILP-LTAD; that is, we can solve the linear programming relaxation and use it to obtain an optimal solution for the MILP in (5).

Lemma 1 *For any* $\mathbf{x}$*, if* $\mathbf{m}_{LP}^* = \mathbf{0}$*, then* $(\mathbf{w}_{LP}^*, \mathbf{m}_{LP}^*)$ *is an optimal solution of MILP-LTAD.*

Proof Let f_{LP}^* and f_{MILP}^* be the optimal solutions of LP-LTAD and MILP-LTAD respectively. If $\mathbf{m}_{LP}^* = \mathbf{0}$ then

$$f_{LP}^* = \sum_{i=1}^{n} \|w_i^* \mathbf{x}_i - \mathbf{m}_{LP}^*\|_1 = \sum_{i=1}^{n} w_i^* \|x_i\|_1 = \sum_{i=1}^{h} \|x_{(i)}\|_1$$

which implies that $w_i^* = 1$ if $i = (i)$ and zero otherwise; or equivalently $w_{LP}^* \in \{0,1\}^n$. Thus, $(\mathbf{w}_{LP}^*, \mathbf{m}_{LP}^*)$ is feasible to MILP-LTAD and $f_{LP}^* = f_{MILP}^*$.

Lemma 1 implies that if we could transform the data in such a way that $\mathbf{m}^*_{LP}$ gets closer to zero, then we can just solve the LP problem to obtain an approximated solution for the LTAD. This leads to the procedure described in Algorithm 4.1, where the data is iteratively transformed such that its median value decreases monotonically until it is less than some tolerance value ϵ.

Algorithm 4.1 Linear Programming Approach for Solving LTAD

```
Input  : data $X_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, coverage $h$, accuracy $\epsilon$
Output: a set of $h$ data points of $X_n$ as indicated by the characteristic vector $\mathbf{w}_{LP}$
while TRUE →
    $(\mathbf{w}^*_{LP}, \mathbf{m}^*_{LP})$ = LP-LTAD$(X_n, h)$
    if $\|\mathbf{m}^*_{LP}\| < \epsilon$ →
        return $\mathbf{w}^*_{LP}$
    else
        $\mathbf{x}_i := \mathbf{x}_i - \mathbf{m}^*_{LP},\ \forall\, i = 1, ..., n$
    end if
end while
```

5 Solution of the LP relaxation

In Algorithm 4.1 we have to solve the associated linear programming problem in each iteration, until $\mathbf{m}_{LP}$ converges to a value smaller than ϵ. The LP as defined in (6), has $(np + n + p)$ decision variables and $(2np + 2n + 1)$ constraints and can be solved efficiently for relatively small (n, p). However, for large values of n, p, e.g. $n = 10000$ and $p = 100$, the problem has a million decision variables and two million constraints. Although this is still solvable, we need to find an efficient solution method since there will be multiple calls of this LP by Algorithm 4.1 to obtain the final solution. In what follows we will exploit the special structure of the problem to develop such a method.

Given $\mathbf{w}$, let $m_j(\mathbf{w})$ be the corresponding median of vector $\{w_i x_{ij} : i = 1, \ldots, n\}$. This means $\mathbf{m}(\mathbf{w}) = (m_1(\mathbf{w}), \ldots, m_p(\mathbf{w}))$ is an optimal solution of (6) for a fixed $\mathbf{w}$. Letting $f(\mathbf{w}) = \sum_{i=1}^n \|w_i \mathbf{x}_i - \mathbf{m}(\mathbf{w})\|_1$ we can write (6) as

$$\begin{aligned} \min_{\mathbf{w}} \quad & f(\mathbf{w}), \\ s.t. \quad & \sum_{i=1}^{n} w_i = h, \\ & 0 \le w_i \le 1. \end{aligned}$$

Here we have transformed the original problem into a new optimization problem in $\mathbb{R}^n$ that has a nice constraint set. However, the objective function is non-linear. Rewriting $f(\mathbf{w})$ we have

$$\begin{aligned} f(\mathbf{w}) &= \sum_{i=1}^{n}\sum_{j=1}^{p} |w_i x_{ij} - m_j(\mathbf{w})| \\ &= \sum_{j=1}^{p}\sum_{i=1}^{n} |w_i x_{ij} - m_j(\mathbf{w})| \\ &= \sum_{j=1}^{p}\sum_{i=1}^{n} |w_i x_{ij} - median(\{w_i x_{ij} : i = 1, \ldots, n\})| \\ &= \sum_{j=1}^{p}\left[\sum_{i=1}^{n} w_i x_{ij} - 2 \min_{|S|=n/2} \sum_{k\in S} w_k x_{ik}\right], \end{aligned}$$

which is a piece-wise convex function because it is the sum of a linear function and the maximum of linear functions. This in turn implies that we can solve the problem using a projected subgradient method, which is shown in Algorithm 5.1.

Algorithm 5.1 Subgradient method for solving LP-LTAD

```
Input  : initial $\mathbf{w}^0$, $\alpha = 1$ and tolerance $\epsilon$
Output: solution $\mathbf{w}^*$ to LP-LTAD
$k = 0$
while TRUE →
    Find subgradient $\mathbf{d}^k = \nabla f(\mathbf{w})$ and set $\bar{\mathbf{w}} = \mathbf{w}^k - \alpha \mathbf{d}^k$
    Find the projection $\mathbf{w}^{k+1}$ of $\bar{\mathbf{w}}$ on the polyhedron $F = \{\mathbf{w} : \sum_{i=1}^{n} w_i = h, 0 \le w_i \le 1\}$
    if $\left\|\mathbf{w}^{k+1} - \mathbf{w}^k\right\| < \epsilon$ →
        return $\mathbf{w}^{k+1}$
    else
        $k = k + 1$
    end if
end while
```

In order to apply the projected subgradient method depicted in Algorithm 5.1, we need to resolve the following three issues: (a) find good initial starting $\mathbf{w}^0$, (b) compute the subgradients, and (c) perform efficient projection onto the polyhedron. Methods for resolving these issues will be presented in the next subsections.

5.1 Finding good initial starting point $\mathbf{w}^0$

We will find an initial starting point $\mathbf{w}^0$ by finding a local optimal solution $(\mathbf{w}, \mathbf{m})$ for the original problem. The idea is to start with an arbitrary initial solution $(\mathbf{w}^0, \mathbf{m}^0)$, set $k = 0$, and repeat the following steps:

(a) Fix $\mathbf{m} = \mathbf{m}^k$ and solve for the corresponding optimal $\mathbf{w}^{k+1}$

(b) If $\left\|\mathbf{w}^{k+1} - \mathbf{w}^k\right\| < \epsilon$, return $\mathbf{w}^{k+1}$ and terminate the procedure. Otherwise, fix $\mathbf{w} = \mathbf{w}^{k+1}$ and solve for the corresponding optimal $\mathbf{m}^{k+1}$. Set $k = k + 1$ and go back to step (a).

In step (b), for each fixed $\mathbf{w}$ finding the corresponding optimal $\mathbf{m}$ is easy, as m_j can be set as the median of $\{w_i x_{ij} : i = 1, \ldots, n\}$. Finding the optimal $\mathbf{w}$ for each fixed $\mathbf{m}$ is non-trivial unless we reformulate it as an LP, but this will be computationally inefficient for large (n, p). A more efficient method is to use a subgradient method to solve the Lagrangian dual problem by noticing that the problem has only one linking constraint $\sum_{i=1}^{n} w_i = k$. Specifically, for a given $\mathbf{m}$, we need to solve

$$\begin{aligned} \min_{\mathbf{w}} \ & \sum_{i=1}^{n} ||w_i \mathbf{x}_i - \mathbf{m}||_1 \ , \\ s.t. \ & \sum_{i=1}^{n} w_i = k, \\ & 0 \le w_i \le 1. \end{aligned} \tag{7}$$

Let δ be the Lagrangian multiplier of the equality constraint (7). The Lagrangian dual problem is

$$\max_{\delta} \left(k\delta + \min_{0 \le w_i \le 1} \ \sum_{i=1}^{n} ||w_i \mathbf{x}_i - \mathbf{m}||_1 - \delta \mathbf{w}^T \boldsymbol{e} \right),$$

which can be further simplified as

$$\max_{\delta} \ \left(k\delta + \sum_{i=1}^{n} \left[\min_{0 \le w_i \le 1} \ |w_i \mathbf{x}_i - \mathbf{m}| - \delta w_i \right] \right).$$

For each fixed δ, the inner problem has a closed form solution for w_i by noticing that the function $g_i(w_i) = ||w_i \mathbf{x}_i - \mu||_1 - \delta w_i$ is piece-wise convex with at most $(p+1)$ pieces that join each other at m_j / x_{ij}. This means we can find the optimal w_i by simply comparing the objective values at those joints that belong to $[0, 1]$. As the inner problem has a closed-form solution and as the outer problem has only a single variable δ, the Lagrangian dual problem can be solved very efficiently. To summarize, we can repeatedly find improving $(\mathbf{w}^k, \mathbf{m}^k)$ and stop the process at a local optimal solution of (6).

For finding subgradients, we notice that

$$f(\mathbf{w}) = \sum_{i=1}^{n} ||w_i \mathbf{x}_i - \mathbf{m}(\mathbf{w})|| = \sum_{j=1}^{p} \underbrace{\sum_{i=1}^{n} |w_i x_{ij} - m_j(\mathbf{w})|}_{f_j(\mathbf{w})},$$

and hence

$$\frac{\partial f}{\partial w_k} = \sum_{j=1}^{p} \frac{\partial f_j}{\partial w_k} = \sum_{j=1}^{p} x_{kj} sign(w_k x_{kj} - m_j(\mathbf{w})).$$

5.2 Finding a projection

The projection of a point on a polyhedron can be found by solving a convex quadratic optimization problem. However, this is not a computationally efficient way and we need to find an alternative by exploiting the special constraint set for $\mathbf{w}$. Notice that this includes only one hyperplane $\sum_{i=1}^{n} w_i = k$ and a set of box constraints. Thus, the projection of any point $\mathbf{w}$ into this polyhedron can be found through two steps:

(a) Finding the projection $\mathbf{w}_P$ of $\mathbf{w}$ onto the plane $\sum_{i=1}^{n} w_i = k$ which has a closed form solution.
(b) Finding the projection $\mathbf{w}_B$ of $\mathbf{w}_P$ into the box constraints. This is simply done by setting:

$$w_{B,i} = \begin{cases} w_{P,i} & \text{if } 0 \le w_{P,i} \le 1, \\ 0 & \text{if } w_{P,i} < 0, \\ 1 & \text{if } w_{P,i} > 1. \end{cases}$$

6 Computational experiments

In this section the performance of LP-LTAD estimator is compared against the performance a heuristic iterative algorithm for solving the LTED estimator based on the Algorithm 2.1 given in [5]. The solutions of the associated problems (6) and (5), were computed using the solver FortMP/QMIP which is a Fortran code provided by Mitra et al. [10]. The computation of the LTED solutions were obtained by a MATLAB implementation of the algorithm in [5].

6.1 Empirical efficiency

Most of the robust estimators in the literature choose a priori a coverage of $h = \frac{n}{2}$, which yields a clean subsample of minimum size. However, if there are fewer outliers in the sample than half of the observations, then information will be discarded when calculating robust estimates based on this. As a consequence these estimates suffer from low efficiency. One solution to this problem is to adapt h, resulting in more efficient estimators which have lower breakdown points. In other words, most robust estimators have to deal with this robustness versus efficiency trade off.

A typical procedure for empirically evaluating the efficiency of robust estimators is to apply the estimators on a clean data set and compare their performance. We conducted a simulation with a a sample data set of 100 observations that follow the standard normal distribution with $N(\mathbf{0}, \mathbf{I})$. After 100 replications, the comparison criterion is the average classical center estimate, or median, for the different coverage sizes $h = 50\%, 60\%, 70\%$ and 80% as it is demonstrated in Table 1. We observe from Table 1, that the proposed robust location estimator LP-LTAD improves significantly with respect to efficiency, as the coverage h decreases. For the smallest coverage $h = 0.5$, which means that 50% of the observations are considered as clean, it yields a location estimate which is the true value $\mathbf{m} = \mathbf{0}$. On the contrary, the ordinary LTED losses in efficiency as the coverage decreases, resulting in a biased location estimate.

coverage h	LP-LTAD	LTED
50%	0.0000	0.0308
60%	0.0001	0.0215
70%	0.0009	0.0210
80%	0.0011	0.0193

Table 1: Median estimate for data set of normal distribution, $N(0, 1)$

6.2 Real data

For our computational experiment involving real data, we used the data by Roelant and Aelst [13] for the L1-type estimator. The data set originates from the *The Data and Story Library* (`http://lib.stat.cmu.edu/DASL/Stories/Forbes500CompaniesSales.html`), which contains facts regarding 79 companies selected from the Forbes 500 list of 1986. We considered the following six variables: *assets* (amount of assets in the company in millions), *sales* (amount of sales in millions), *market-value* (market-value of the company in millions), *profits* (profits in millions), *cash-flow* (cash-flow in millions) and *employees* (number of employees in thousands). We applied the L1-type, LP-LTAD and LTED estimators to find an estimate of location. Table 2 compares the location estimates with the empirical mean. Clearly, there are large differences between the locations estimates of the above estimators with the empirical mean. The empirical means are much higher than all estimators. These differences are caused by the presence of outliers in the data set, which greatly affect the mean statistic. We can see that all estimators perform comparably with respect to obtaining a robust location estimate.

	Assets	Sales	Market-value	Profits	Cash-flow	Employees
Empirical mean	5940.53	4178.29	3269.75	209.84	400.93	37.60
L1-type mean	2679.33	1757.50	1099.14	89.90	164.10	15.63
LP-LTAD	2677.21	1752.72	1096.78	88.71	164.02	15.19
LTED	2679.15	1753.10	1097.12	89.01	164.09	15.94

Table 2: Estimate for the location of the Forbes data set

6.3 Simulation results

To study the finite-sample robustness and efficiency of the three robust location estimates, we performed simulations with contaminated data sets. In each simulation we generate 100 data sets based on a multivariate normal distribution $N_p(\mathbf{0}, \mathbf{I})$ with $p = 1, 2, 3, 5$ and sample sizes $n = 50, 100$. To generate contaminated data sets we replaced $\epsilon \in \{20\%, 40\%\}$ of the data $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ with outliers from a multivariate normal distribution, $N_p(3.3, 0.3^2)$. For each of the aforementioned data sets we obtained the LP-LTAD local estimate $\hat{\mathbf{m}}$ by computing the solution of the problems formulated in (6). After 100 replications, we recorded the mean square errors (MSE) as a performance criterion for comparison between the robust local estimates, given as

$$\text{MSE} = \frac{\sum_{i=1}^{100} ||\hat{\mathbf{m}}||^2}{100}.$$

Moreover, we also recorded the computational time in CPU seconds for each method which is displayed in parenthesis next to the MSE in the tables that follow.

The results are shown in Tables 3, 4 and 5. In Table 3, the two estimators use half sample coverage $h = 50\%$. We observe that the performance of the new approach LP-LTAD is quite competitive compared to the LTED. In Table 4 the proposed model LP-LTAD uses as coverage $h = 20\%$, while the LTAD estimator $h = 50\%$ which is the smallest that we can use in this case. Note that the new estimator LP-LTAD outperforms the LTED in all instances. The results in Table 5 reveal that when the sample size increases all the estimators have similar performance.

		LP-LTAD	LTED
	ϵ	50%	50%
$p = 1$	0%	0.0009 (0.01)	0.0007 (0.01)
	20%	0.0015 (0.01)	0.0012 (0.01)
	40%	0.0069 (0.01)	0.0071 (0.01)
$p = 2$	0%	0.0010 (0.01)	0.0013 (0.01)
	20%	0.0300 (0.01)	0.0295 (0.01)
	40%	0.0387 (0.01)	0.0314 (0.01)
$p = 3$	0%	0.0091 (0.01)	0.0020 (0.01)
	20%	0.0410 (0.01)	0.0791 (0.01)
	40%	0.0415 (0.01)	0.0834 (0.01)

Table 3: MSE of local estimates, $n = 50$, $p = 1, 2, 3$

In order to investigate the effect of the presence of a correlation structure within the simulated data on the performance of the algorithms, we used a covariance matrix P' for data generation, with elements 1 in the diagonal, and numbers ρ as off-diagonal elements. We chose a value of $\rho = 0.70$ among the different simulation scenarios. Similar structures for simulated correlation data have been proposed by Maronna and Zamar [9] and Hubert et al. [8]. The results are shown in Tables 6, 7 and 8. We observe that the effect of the correlation does not influence the performance of the LP-LTAD estimator. It should be noted that the reduction of coverage from 50% to 20% does not improve the performance of the LP-LTAD.

		LP-LTAD	LTED
	ϵ	20%	50%
$p = 1$	0%	0.0001 (0.01)	0.0007 (0.01)
	20%	0.0006 (0.01)	0.0012 (0.01)
	40%	0.0039 (0.01)	0.0071 (0.01)
$p = 2$	0%	0.0009 (0.01)	0.0013 (0.01)
	20%	0.0244 (0.01)	0.0295 (0.01)
	40%	0.0281 (0.01)	0.0314 (0.01)
$p = 3$	0%	0.0017 (0.01)	0.0020 (0.01)
	20%	0.0101 (0.01)	0.0791 (0.01)
	40%	0.0105 (0.01)	0.0834 (0.01)

Table 4: MSE of local estimates, $n = 50$, $p = 1, 2, 3$

		LP-LTAD	LTED
	ϵ	20%	50%
$p = 1$	0%	0.0001 (0.01)	0.0006 (0.01)
	20%	0.0004 (0.01)	0.0011 (0.01)
	40%	0.0024 (0.01)	0.0028 (0.01)
$p = 3$	0%	0.0008 (0.01)	0.0012 (0.01)
	20%	0.0081 (0.01)	0.0601 (0.01)
	40%	0.0151 (0.01)	0.0714 (0.01)
$p = 5$	0%	0.0015 (0.02)	0.0061 (0.02)
	20%	0.0094 (0.02)	0.0715 (0.02)
	40%	0.0171 (0.02)	0.0924 (0.02)

Table 5: MSE of local estimates, $n = 100$, $p = 1, 3, 5$

		LP-LTAD	LTED
	ϵ	50%	50%
$p = 2$	0%	0.0005 (0.01)	0.0022 (0.01)
	20%	0.0041 (0.01)	0.0118 (0.01)
	40%	0.0376 (0.01)	0.1010 (0.01)
$p = 3$	0%	0.0009 (0.01)	0.0151 (0.01)
	20%	0.0213 (0.01)	0.3211 (0.01)
	40%	0.0773 (0.01)	0.3912 (0.01)

Table 6: MSE of local estimates, $n = 50$, correlation ρ=0.7, $p = 2, 3$

		LP-LTAD	LTED
	ϵ	20%	50%
$p = 2$	0%	0.0004 (0.01)	0.0022 (0.01)
	20%	0.0030 (0.01)	0.0118 (0.01)
	40%	0.0164 (0.01)	0.1010 (0.01)
$p = 3$	0%	0.0005 (0.01)	0.0151 (0.01)
	20%	0.0056 (0.01)	0.3211 (0.01)
	40%	0.0225 (0.01)	0.3912 (0.01)

Table 7: MSE of local estimates, $n = 50$, correlation ρ=0.7, $p = 2, 3$

To illustrate the performance of the estimators on data contaminated with *intermediate* outliers, we replaced 20% or 40% of the first rows of the multivariate sample, $N_p(\mathbf{0}, \mathbf{I})$, with intermediate outliers from a multivariate normal distribution $N_p(0.75, 0.5)$, as suggested by Roelant et al. [14]. We prefered to reduce the coverage to $h = 20\%$, among the different scenarios $30\%, 40\%, 50\%$, which enables the new LP approach to identify the intermediate outliers. In the results given by Tables 9 and 10 it is evident that the LP-LTAD outperforms the LTED. This is especially true for heavily contaminated data.

		LP-LTAD	LTED
	ϵ	20%	50%
$p = 3$	0%	0.0009 (0.01)	0.0073 (0.01)
	20%	0.0110 (0.01)	0.0814 (0.01)
	40%	0.0215 (0.01)	0.1011 (0.01)
$p = 5$	0%	0.0019 (0.02)	0.0094 (0.02)
	20%	0.0201 (0.02)	0.1001 (0.02)
	40%	0.0274 (0.02)	0.1703 (0.02)

Table 8: MSE of local estimates, $n = 100$, correlation $\rho = 0.7$, $p = 3, 5$

		LP-LTAD	LTED
	ϵ	20%	50%
$p = 1$	0%	0.0001 (0.01)	0.0006 (0.01)
	20%	0.0001 (0.01)	0.0314 (0.01)
	40%	0.0018 (0.01)	0.1123 (0.01)
$p = 2$	0%	0.0006 (0.01)	0.0244 (0.01)
	20%	0.0391 (0.01)	0.2111 (0.01)
	40%	0.0415 (0.01)	0.2291 (0.01)
$p = 3$	0%	0.0184 (0.01)	0.0810 (0.01)
	20%	0.1120 (0.01)	0.2415 (0.01)
	40%	0.1230 (0.01)	0.2581 (0.01)

Table 9: MSE of local estimates, $n = 50$, $p = 1, 2, 3$

		LP-LTAD	LTED
	ϵ	20%	50%
$p = 1$	0%	0.0001 (0.02)	0.0001 (0.03)
	20%	0.0002 (0.02)	0.0236 (0.03)
	40%	0.0009 (0.02)	0.1123 (0.03)
$p = 3$	0%	0.0009 (0.02)	0.0094 (0.03)
	20%	0.0315 (0.02)	0.2012 (0.03)
	40%	0.0318 (0.02)	0.2094 (0.03)
$p = 5$	0%	0.0009 (0.02)	0.0601 (0.03)
	20%	0.0517 (0.02)	0.2151 (0.03)
	40%	0.0518 (0.02)	0.2204 (0.03)

Table 10: MSE of local estimates, $n = 100$, $p = 1, 3, 5$

Finally we generated a large data set with $n = 500$ and $p = 10, 20$ with the same contamination and distributions as the previous simulations. The results are shown in Tables 11 and 12.

		LP-LTAD	LTED
	ϵ	20%	50%
$p = 10$	0%	0.0098 (0.03)	0.0099 (0.31)
	20%	0.0109 (0.03)	0.0871 (0.31)
	40%	0.0109 (0.03)	0.0882 (0.31)
$p = 20$	0%	0.0104 (0.03)	0.0121 (0.55)
	20%	0.0121 (0.03)	0.1001 (0.55)
	40%	0.0124 (0.03)	0.1012 (0.55)

Table 11: MSE of local estimates, $n = 500$, $p = 10$

In summary, based on the computational results we can conclude that there are negligible differences between LP-LTAD and LTED for non-correlated data and contaminated with strong outliers. In the case of correlated data LP-LTAD has the best performance. Also, if the data is contaminated with intermediate

		LP-LTAD	LTED
	ϵ	20%	50%
	0%	0.0091 (0.03)	0.0098 (0.31)
$p = 10$	20%	0.0108 (0.03)	0.2456 (0.31)
	40%	0.0108 (0.03)	0.2551 (0.31)
	0%	0.0101 (0.03)	0.0099 (0.55)
$p = 20$	20%	0.0131 (0.03)	0.3210 (0.55)
	40%	0.0132 (0.03)	0.3481 (0.55)

Table 12: MSE of local estimates, $n = 500$, $p = 20$

outliers the LP-LTAD is superior because it can work with coverage less than 50% while, on the other hand, the LTED includes some of the intermediate outliers into the coverage set so they become masked. Finally, with respect to the computational time both estimators are comparable for instance sizes up to $n = 100$, but the LP-LTAD is marginally faster than the LTED for larger size instances.

7 Conclusions

In this work, we develop numerical methods for computing the multivariate LTAD estimator based on the L^1 norm, by reformulating its original mixed integer nonlinear formulation. We show that the MINLP is equivalent to an MILP and subsequently to an LP under some conditions on the location estimate. An LP-based iterative approach is then developed for computing the estimator, by transforming the data and solving the resulting linear programs by subgradient optimization. The new LP-LTAD formulation can also be viewed as a new trimming procedure that trims away large residuals implicitly by shrinking the associated observations to zero. The new approach yields a robust location estimate without loosing efficiency. We perform numerical experiments and show that the new estimate performs well even in the case of contaminated and correlated multivariate data. The LP-LTAD procedure can be used when the data involves both type of outliers, strong and intermediate, and also when the coverage is smaller than half the sample observations.

References

1. Charu C. Aggarwal. Outlier detection in graphs and networks. In *Outlier Analysis*, pages 343–371. Springer New York, 2013.
2. Charu C. Aggarwal, Yuchen Zhao and Philip S. Yu. Outlier Detection in Graph Streams. In *Proceedings of the 2011 IEEE 27th International Conference on Data Engineering*, pages 399–409. IEEE Computer Society, 2011.
3. G.W. Bassett. Equivariant, monotonic, 50% breakdown estimators. *The American Statistician*, 45(2):135–137, 1991.
4. Biman Chakraborty, Probal Chaudhuri, and Hannu Oja. Operating transformation retransformation on spatial median and angle test. *Statistica Sinica*, 8:767–784, 1998.
5. Christos Chatzinakos, Leonidas Pitsoulis, and George Zioutas. Optimization techniques for robust multivariate location and scatter estimation *Journal of Combinatorial Optimization*, online, 2015.
6. Peter Filzmoser, Ricardo Maronna, and Mark Werner. Outlier identification in high dimensions. *Computational Statistics & Data Analysis*, 52(3):1694 – 1711, 2008.
7. Thomas P. Hettmansperger. A practical affine equivariant multivariate median. *Biometrika*, 89(4): 851–860, 2002.
8. Mia Hubert, Peter J. Rousseeuw, and Tim Verdonck. A deterministic algorithm for the mcd, 2010.
9. Ricardo A Maronna and Ruben H Zamar. Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, 44(4):307–317, 2002.

10. G. Mitra, M. Guertler, and F. Ellison. Algorithms for the solution of large-scale quadratic mixed integer programming (qmip) models. In *International Symposium in Mathematical Programming*, 2003.
11. N.M. Neykov, P. Čížek, P. Filzmoser, and P.N. Neytchev. The least trimmed quantile regression. *Computational Statistics & Data Analysis*, 56(6):1757–1770, 2012.
12. Hannu Oja. Descriptive statistics for multivariate distributions. *Statistics & Probability Letters*, 1(6): 327–332, 1983.
13. Ella Roelant and StefanVan Aelst. An l1-type estimator of multivariate location and shape. *Statistical Methods and Applications*, 15(3):381–393, 2007.
14. Ella Roelant, Stefan Aelst, and Gert Willems. The minimum weighted covariance determinant estimator. *Metrika*, 70(2):177–204, 2009.
15. Peter J. Rousseeuw. Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*, B:283–297, 1985.
16. Peter J. Rousseeuw and Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223, 1998.
17. P. J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79:871–881, 1984.
18. Mara Tableman. The asymptotics of the least trimmed absolute deviations (LTAD) estimator. *Statistics & Probability Letters*, 19(5):387–398, 1994.