2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Tilted Least-Squares Parameter Estimation of Linear Regression Models in the Presence of Outliers

Biqiang Mu, Er-Wei Bai, and Wei Xing Zheng

Abstract— **The least squares estimator is the most popular identification method. In the absence of prior knowledge on the unknown noise, uniform weights on all samples are often assumed. In reality, potentially unknown contamination is always present and the uniform weights are not necessarily the best. Further, explicit information about the nature of contamination is usually absent. To this end, a relaxed-tilted least squares method is proposed here to assign unequal weights so that the effect of undesired noise contamination can be mitigated. The relaxed-tilted least squares method tilts the uniform prior on the samples so as to move the uniform distribution in a direction that enjoys the smallest estimation error in the neighborhood of the uniform distribution. Theoretical results are established including the ability of outlier removal and the guaranteed parameter convergence in the presence of outliers. Numerical algorithms are proposed and simulated, which support the theoretical derivations.**

Index Terms— **Robust least squares, Outliers, Heavy-tailed noises, System identification, Parameter estimation**

I. Introduction

The least squares (LS) estimator [1]–[5] is the most popular method in system identification. Two of the most appealing properties of the LS estimator are its computational simplicity and no prior knowledge requirement on the unknown noise. In the absence of explicit prior information about the nature of noise, a uniform weight is ostensibly assumed for most of applications. Clearly, the robustness performance of the LS estimator is in question if the noise sequence is so that an unequal weight is called for [6], [7]. There exist many such cases in the literature and some examples are provided here.

1. Outliers: The LS method performs very poorly in the presence of outliers since a single bad data point can make the estimate arbitrarily bad. If these outliers could be identified and removed in identification automatically, then by setting weights to be equal to zero at the outlier data points, the performance of the LS method would be much improved.

This work was supported in part by the National Key R&D Program of China under Grant No. 2022YFA1004700, NIHR15AG061755-01 and NIHR42CA195819.

Biqiang Mu is with the Key Laboratory of Systems and Control, Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China (bqmu@amss.ac.cn).

Er-Wei Bai is with the Department of Electrical and Computer Engineering, University of Iowa, Iowa City, Iowa 52242, USA (er-wei-bai@uiowa.edu).

Wei Xing Zheng is with the School of Computer, Data and Mathematical Sciences, Western Sydney University, Sydney, NSW 2751, Australia (w.zheng@westersydey.edu.au).

2. Heavy-tailed noises, even independent and identically distributed (i.i.d.): It is well known that if the noise distribution is heavy-tailed, then the distribution of the LS estimate inherits a similar heavy-tailed behavior [8]. Therefore, it would make perfect sense to ignore a small portion of data points that are likely corrupted by large magnitude noise or to assign smaller weights at these data points.
3. Uncertainties that could act like outliers. Examples include
   - Independent but non-identically distributed noise: The level of the noise may depend on many factors, so at some data points, the data could be affected differently. This implies that some data points are more reliable or less reliable than others. One certainly wants to ignore those less reliable points or to use them less in identification.
   - Uncertainties due to reduced-order model, multiplicative noise and small nonlinearities: In such cases, the "noise" may not be generated by measurement errors but it depends on the output, the input and nonlinearities. These kinds of noises can be large occasionally and behave like outliers.

Robust estimation aims to handle the cases above such that the estimate is insensitive to noises. One scheme is based on the residuals, which includes the least median squares (LMS) estimator minimizing the median of the squared residuals [9], the least trimmed squares (LTS) estimator minimizing the sum of squared residues over a subset of the given data [10], and the lease absolute deviation (LAD) estimator minimizing the sum of the absolute values of the residuals [11], [12]. All of these methods are proved to be robust against outliers. Both the LMS and LTS estimators involve a combinatorial computational complexity, but the LAD problem can be solved by linear programming. Meanwhile, the LAD estimator is unstable and possibly gives multiple solutions. The idea of the few violated constraints (FVC) method [1] is also very effective in dealing with outliers, but its problem is again a high computational complexity. Recently, compressed sensing techniques were developed to handle outliers in [13] because outliers are usually sparse. However, this technique assumes some restricted isometry properties that do not hold in most of identification settings.

In this paper, we propose a parametric way to assign unequal weights to each sample so that the effect of undesired noise contamination can be mitigated. The scheme is to tilt a uniform prior on the samples so as to move the uniform

distribution in a direction that enjoys the smallest estimation error in the neighborhood of the uniform distribution. The idea of "tilting" is not new and was proposed in [14]–[16]. Based on the maximum likelihood criterion, the approach in [15] tilts a likelihood function to enhance the robustness. But the resultant computation is highly nonlinear, which is solved by grid search or Newton-Raphson numerical methods. Since it works on the likelihood function explicitly, the distributions function of the unknown noise needs to be available. The idea of tilting can also be found in [14], [16] in a risk minimization framework. It does not however work on the least squares criterion; instead it finds the minimum bound on the risk minimization.

Moreover, we relax the nonlinear constraint presented in [15] to a linear constraint that makes computation much simpler and we call the corresponding estimator the relaxed-tilted least squares (RTLS) estimator. Further, the proposed RTLS estimator works directly on the sampled data and no distribution information on the unknown noise is needed. Therefore, the proposed algorithm enhances the robustness of the LS estimator in the presence of unknown noise contamination and at the same time maintains two appealing properties of the LS estimator: simplicity and no requirement on the unknown noise. The main contributions of this paper are to show that the proposed RTLS estimator is able to remove unknown outliers. Finally, it is noteworthy that the weighted least squares (WLS) estimator also allows unequal weights but the weights are pre-calculated. In contrast, the weights in the proposed RTLS estimator are adjusted based on the data as a part of the estimator.

The layout of the paper is as follows. The relaxed-tilted least squares estimator is proposed in Section II along with some preliminary properties. Section III focuses on the ability of the RTLS estimator in removing unknown outliers. Section IV addresses computation issues and develops numerical algorithms of the RTLS estimator. The algorithm for implementing the RTLS estimator is numerically extensively tested and simulations results are reported in Section V. Finally, some concluding remarks are drawn in Section VI. Note that the proofs of the lemmas and theorems of this paper are omitted due to limited space.

II. Relaxed-Tilted Least Squares Estimators and Preliminary Analysis

Consider the finite impulse response (FIR) system described by

$$y_k = x_k^T \beta^* + v_k, \quad k = 1, 2, \cdots, n, \tag{1a}$$

$$x_k = (u_{k-1}, \cdots, u_{k-p})^T, \tag{1b}$$

where β^* is the unknown true parameter vector of dimension $p > 0$, and $x_k \in \mathbb{R}^p$, $y_k \in \mathbb{R}$ and $v_k \in \mathbb{R}$ are the regressor, the output and the disturbance at time k, respectively. By stacking the data in the way that $Y = (y_1, \cdots, y_n)^T$, $X = (x_1, \cdots, x_n)^T$, and $V = (v_1, \cdots, v_n)^T$, a standard linear regression model is constructed as below:

$$Y = X\beta^* + V. \tag{2}$$

A. Ordinary Least Squares Estimator

The ordinary least squares (OLS) estimator, minimizing the average of the squared residuals

$$\frac{1}{n}\sum_{k=1}^{n} r_k^2(\beta), \tag{3}$$

where $r_k(\beta) \triangleq y_k - x_k^T\beta$ is the residual of the k-th observation for a given β, is a prevalent estimation method for inferring the parameter vector β^*.

The OLS estimator is optimal among all the unbiased estimators if the disturbance v_k is i.i.d., but it might not be optimal or further might deteriorate if the disturbance is heteroscedastic or has large magnitudes from different sources, called outliers. To measure how sensitive an estimator is to outliers (e.g., arbitrarily large observations), we use the concept called the breakdown point of an estimator. The breakdown point of an estimator is the proportion of outliers that an estimator can handle before giving an arbitrarily large error [17]. Mathematically, the breakdown point can be defined as [17]

$$B(X,Y) = \frac{1}{n}\min_{k\in\{1,2,\cdots,n\}}\Big\{k \,\Big|\, \sup_{Q_{n,k}} \|\widehat{\beta}_{Q_{n,k}} - \beta^*\| = \infty\Big\},$$

where $Q_{n,k}$ is the empirical distribution of n data points in which at least k sample points out of $\{X, Y\}$ are replaced and $\widehat{\beta}_{Q_{n,k}}$ is the corresponding estimate. The OLS estimator is known to be sensitive to arbitrarily large observations and its breakdown point is $\frac{1}{n}$, that is, only one single arbitrarily large observation can make the OLS estimate arbitrarily unreliable.

B. Ordinary Tilted Least Squares Estimators

To guarantee an estimator to be robust, another scheme falls into the weighted least squares (WLS) framework, which introduces a weight for each squared residual to reflect the significance of data points for estimation. That is, the loss function of the WLS estimator is given by

$$J_n(\beta, w) \triangleq \sum_{k=1}^{n} w_k r_k^2(\beta), \tag{4}$$

where $w \triangleq (w_1, w_2, \cdots, w_n)^T$ is the weight vector. The developed robust estimator should automatically detect outliers by assigning different weights estimated from the noise-contaminated data. For it to make sense, the weight vector has to satisfy some reasonable restrictions expressed by the set

$$\mathcal{W} \triangleq \big\{w \mid \mathbf{0} \preccurlyeq w \preccurlyeq \mathbf{1}, \mathbf{1}^T w = 1\big\}, \tag{5}$$

where $a \preccurlyeq b$ means the pointwise inequality, and $\mathbf{0}$ and $\mathbf{1}$ are the column vectors with all elements being 0 and 1, respectively. For the squared residuals $r_k^2(\beta), 1 \le k \le n$, let us denote its ordered form in the ascending order by

$$r_{[1]}^2(\beta) \le r_{[2]}^2(\beta) \le \cdots \le r_{[n]}^2(\beta).$$

We hope that the estimated weight vector has the property:

The weights of outliers will be assigned zero or very close to zero such that the influence of the outliers on the estimate can be totally removed or cured.

It is obvious that the following problem

$$\min_{\beta\in\mathbb{R}^p, w\in\mathcal{W}} J_n(\beta, w) \tag{6}$$

is a poorly designed optimization problem for achieving the goal. Because there always exists some β such that $y_1 = x_1^T\beta$, the loss function $J_n(\beta, w)$ can reach its minimum 0 at the weight vector $w = (1, 0, \cdots, 0)^T$. However, this weight vector is not what we want.

To make the idea work along this line, more constraints on the weight vector w need to be imposed. We now introduce the Kullback-Leibler (KL) divergence [18] constraint $\mathcal{D}(w)$, where

$$\mathcal{D}(w) \triangleq \sum_{k=1}^{n} w_k \ln\frac{w_k}{1/n} = \sum_{k=1}^{n} w_k \ln(nw_k)$$

is the KL divergence deviation of w from the discrete uniform distribution $w^\dagger \triangleq (\frac{1}{n}, \cdots, \frac{1}{n})^T$. The variable $\delta > 0$ is adopted to control the amount of the allowable deviation. Now following the idea of tilting [15], the ordinary tilted least squares (OTLS) estimator can be defined as

$$\left[\widehat{\beta}_{n,\delta}^{otls}, \widehat{w}_{n,\delta}^{otls}\right] \triangleq \underset{\beta\in\mathbb{R}^p, w\in\mathcal{W}, \mathcal{D}(w)\le\delta}{\arg\min} J_n(\beta, w). \tag{7}$$

The idea of the OTLS estimator is to tilt the uniform prior on the data points so as to move the uniform distribution in a direction that enjoys the smallest estimation error in the δ-neighborhood of the uniform distribution defined by the KL divergence. Therefore, the possible robust properties of the OTLS estimator are determined by the nonlinear constraint $\mathcal{D}(w) \le \delta$. Meanwhile, the nonlinear constraint is also the source of the difficulty for clearly exploring the OTLS estimator. In particular, we are interested in the following problems:

1. how does the weight estimate $\widehat{w}_{n,\delta}^{otls}$ depend on the squared residuals $r_k^2(\beta)$?
2. further how many zero elements does $\widehat{w}_{n,\delta}^{otls}$ have exactly if $\widehat{w}_{n,\delta}^{otls}$ includes the zero element?

C. Relaxed-Tilted Least Squares Estimators

Lemma 1: The KL divergence $\mathcal{D}(w)$ satisfies the property

$$0 \le \mathcal{D}(w) \le \ln(n)$$

for all $w \in \mathcal{W}$.

Let us consider a family of disjoint subsets of $\mathcal{W}$ defined by

$$\mathcal{W}^m \triangleq \left\{w \in \mathcal{W} \mid \sharp(w = 0) = m\right\},\ m = 0, \cdots, n \tag{8}$$

with $\sharp(w = 0)$ denoting the number of zero elements of the weight vector w. It is clear that

$$\mathcal{W}^l \ne \mathcal{W}^m \ \text{ for } l \ne m \text{ and } \bigcup_{j=0}^{n-1} \mathcal{W}^j = \mathcal{W}.$$

The KL divergence constraint $\mathcal{D}(w) \le \delta$ on the subsets $\mathcal{W}^j$ displays an attractive property as indicated below.

Lemma 2: For all $w \in \mathcal{W}^m$, there holds that

1) $\ln\frac{n}{n-m} \le \mathcal{D}(w) < \ln(n)$;
2) $\mathcal{D}(w)$ over $\mathcal{W}^m$ attains its unique minimum value $\ln(n/(n-m))$ at the points with m zero elements and other elements being $1/(n-m)$.

Motivated by Lemma 2, a meaningful question naturally arises: whether does there exist other estimator that can approximate the OTLS estimator with the constraint $\mathcal{D}(w) \le \delta$ taking the special values

$$\delta_m \triangleq \ln\Big(\frac{n}{n-m}\Big),\ m = 0, 1, 2, \cdots, n-1$$

such that some elements of its weight vector take the exact zero?

To this end, we define the relaxed-tilted least squares (RTLS) estimator

$$\left[\widehat{\beta}_{n,m}^{rtls}, \widehat{w}_{n,m}^{rtls}\right] \triangleq \underset{\beta\in\mathbb{R}^p, w\in\mathcal{W}'_m}{\arg\min} J_n(\beta, w), \tag{9a}$$

$$\mathcal{W}'_m \triangleq \left\{w \,\middle|\, \mathbf{0} \preccurlyeq w \preccurlyeq \frac{1}{n-m}\mathbf{1}, \mathbf{1}^T w = 1\right\} \tag{9b}$$

for some $m = 0, 1, 2, \cdots, n-1$. Then the following theorem will provide an affirmative answer to the above question.

Theorem 1: There holds that

$$\widehat{w}_{n,m}^{rtls} \in \mathcal{W}^m,\ \mathcal{D}(w_{n,m}^{rtls}) = \delta_m,$$
$$J_n(\widehat{\beta}_{n,\delta_m}^{otls}, \widehat{w}_{n,\delta_m}^{otls}) \le J_n(\widehat{\beta}_{n,m}^{rtls}, \widehat{w}_{n,m}^{rtls}).$$

for all $m = 0, 1, 2, \cdots, n-1$.

The weights $\widehat{w}_{n,m}^{rtls}$ have the sparsity and satisfies the nonlinear constraint $\mathcal{D}(w) \le \delta_m$. Theorem 1 states that the weights produced by the RTLS estimator return m exactly zero elements and further the loss function of the RTLS estimator is an upper bound of that of the OTLS estimator.

D. Connection to Linear Trimmed Squares Estimator

The conclusion that all of the l nonzero elements of $\widehat{w}_{n,m}^{rtls}$ are equal to $1/l$ implies that the RTLS estimator searches for the parameter vector β that can minimize the l smallest squared residuals. Actually, the least trimmed squares (LTS) estimator developed in [10] is defined by directly minimizing the l smallest squared residuals

$$\widehat{\beta}_{n,m}^{lts} \triangleq \underset{\beta\in\mathbb{R}^p}{\arg\min} K_n(\beta),\quad K_n(\beta) \triangleq \frac{1}{l}\sum_{k=1}^{l} r_{[k]}^2(\beta). \tag{10}$$

Based on the ideas of the RTLS and LTS estimators, it is reasonably expected that there should be some connection between them.

We first present the following assumption on the regressors and the noise.

Assumption 1:

1. The remaining regressors by removing any $\{x_{i_1}, x_{i_2}, \cdots, x_{i_m}\}$ from $\{x_k,\ k = 1, 2, \cdots, n\}$ are persistently exciting (PE), i.e.,

$$d_1 I \le \frac{1}{l}\sum_{k\ne i_1,\cdots,i_m} x_k x_k^T \le d_2 I,\quad l \triangleq n - m$$

for some $d_2 > d_1 > 0$.

2. The noise vector V is not a linear combination of the columns of the matrix X.

These two conditions are necessary. Without the PE condition, the uniqueness of the estimate is not guaranteed. For Assumption 1.2, suppose $V = X\gamma$, then $Y = X\beta^* + V = X(\beta^* + \gamma)$. This implies that the parameter vector β^* is not identifiable. In reality, if v_k is random, then this situation does not happen with probability one.

Thus, the following proposition establishes their relation.

Proposition 1: Under Assumption 1, there holds that

$$J_n\big(\widehat{\beta}^{rtls}_{n,m}, \widehat{w}^{rtls}_{n,m}\big) = K_n\big(\widehat{\beta}^{lts}_{n,m}\big), \tag{11a}$$

$$\widehat{\beta}^{rtls}_{n,m} = \widehat{\beta}^{lts}_{n,m}. \tag{11b}$$

Directly solving the LTS problem (10) involves running $\binom{n}{m}$ LS estimates, which has a combinatorial computational complexity. To establish a direct connection between the RTLS and LTS estimators, we rewrite the LTS estimator in the following form:

$$\big[\widehat{\beta}^{lts}_{n,m}, \widehat{w}^{lts}_{n,m}\big] \triangleq \underset{\beta\in\mathbb{R}^p, w\in\mathcal{W}''_m}{\arg\min} J_n(\beta, w), \tag{12a}$$

$$\mathcal{W}''_m \triangleq \big\{w \mid w_k = \{0, 1/l\}, \mathbf{1}^T w = 1\big\}. \tag{12b}$$

It is clear that the redefined form (12) is equivalent to the original LTS estimator (10). The difference between the RTLS and LTS estimators is caused by two different constraints imposed on the weight vector w. In the LTS estimator, the constraint is $w_k = \{0, \frac{1}{l}\}, \sum_{k=1}^n w_k = 1$, which is an integer programming problem, while the constraint of the RTLS estimator is $0 \le w_k \le \frac{1}{l}$, $\sum_{k=1}^n w_k = 1$, which is a linear constraint and makes calculation much easier.

E. Breakdown Point

We now study the breakdown point of the RTLS estimator.

Lemma 3: Suppose that Assumption 1 holds and an upper bound m on the number of outliers of the n data points $\{x_k, y_k\}$ is available. Then the breakdown point of the RTLS estimator (9) is $(m+1)/n$.

It is interesting and surprising to observe that the breakdown point of the RTLS estimator (9) is $(m+1)/n$ which could be higher than 50%. It is well known in the literature that 50% is the highest breakdown point that any estimator can achieve. For example, the breakdown point of the LMS estimator is 50%. Note, however, that in (9), the upper bound m is assumed to be known, which is not assumed in many robust estimators. The high breakdown point of the RTLS estimator (9) is achieved only when the prior information on the upper bound m of the number of outliers becomes available and the remaining regressors are persistently exciting by Assumption 1.

III. Removal of Outlier by RTLS

This section aims at exploring conditions under which the RTLS estimator can remove outliers for different settings. In detail, here we assume that the upper bound m on the number of unknown outliers is available. We expect that a set of m data points, which contains all possible outliers, can be detected and removed by using the RTLS estimator (9) if the amplitudes of outliers are large.

A. Finite Sample Size with $n > m$

In this subsection, we focus on the finite sample case and plan to prove that the data points corresponding to the outliers with large amplitude will be removed by the RTLS estimator. Since the RTLS method actually solves the LTS problem (10), there are exactly l data points that are selected to compute the estimate of the parameter vector β.

Let us denote the indexes of the normal part and the outliers of the data by $\mathcal{N}$ and $\mathcal{O}$, respectively. Clearly, we have $\mathcal{N} \cup \mathcal{O} = \{1, 2, , \cdots, n\}$ and $\mathcal{N} \cap \mathcal{O} = \emptyset$. For convenience, we also denote the indexes of the data selected by the RTLS estimator by $\mathcal{R}$. Moreover, we use $Y_{\mathcal{S}}$, $X_{\mathcal{S}}$, and $V_{\mathcal{S}}$ to denote the vectors and matrices consisting of the vectors and matrices Y, X, and V with the indexes being S, where S can be $\mathcal{N}$, $\mathcal{O}$, or $\mathcal{R}$. In particular, when $\mathcal{N} \cup \mathcal{R} \neq \emptyset$, without loss of generality, we put the same indexes included in both $\mathcal{N}$ and $\mathcal{R}$ at the same positions in the vectors and matrices. Denote the singular value decomposition (SVD) of $X_{\mathcal{R}}$ by

$$X_{\mathcal{R}} = U \begin{pmatrix} S \\ 0 \end{pmatrix} Q^T = \begin{pmatrix} U_1 & U_{12} \\ U_{21} & U_2 \end{pmatrix} \begin{pmatrix} S \\ 0 \end{pmatrix} Q^T.$$

Assumption 2: There exist some constants $\delta_i > 0$, $i = 1, 2, 3, 4$ so that

$$\delta_1 I \le U_2 U_2^T \le \delta_2 I, \quad \delta_3 I \le \frac{X_{\mathcal{N}}^T X_{\mathcal{N}}}{l}, \quad \frac{X_{\mathcal{R}}^T X_{\mathcal{R}}}{l} \le \delta_4 I.$$

Theorem 2: Consider the RTLS estimator (9) and suppose that Assumptions 1 and 2 hold. Let $c_1 = \max_{i\in\mathcal{N}} |v_i|$. Then there exists a constant $c_2 > 0$, that could be a function of c_1, n and m, such that $i \notin \mathcal{R}$ if $|v_i| > c_2$.

In other words, Theorem 2 states that any outlier can be identified and removed by the RTLS estimator as long as the outlier is large.

We make some comments here.

- The normal part of the noise can be deterministic or random. Moreover, no assumption or any property on possible outliers is imposed.
- If the noise is deterministic, then the bound c_1 is well defined. If the normal part of the noise is stochastic and has a bounded support, e.g., uniform distribution, then c_1 is also well defined. Even for noises of an unbounded support, e.g., Gaussian noise of zero mean and unit variance with $n = 1000$, $c_1 \le 5$ is well defined with a high probability > 0.9997. Furthermore, this probability can be made arbitrarily high if c_1 is enlarged.
- No knowledge of c_1 is needed or assumed in the derivation.

B. Asymptotic Results as $n \to \infty$

The results in the preceding subsection hold for all fixed $n > m$. Now we consider an asymptotic case when $n \to \infty$. Intuitively, the number of possible outliers increases as the total data length $n \to \infty$. The question is how to identify

and remove these outliers. The problem is of course that m is unknown and the outliers are unknown. To this end, it is reasonable to assume that the upper bound $m = \alpha n$ on the unknown outliers is available. For instance, the maximum number of outliers is no more than 5% or 10% of the total data points, i.e., $m = 0.05n$ or $m = 0.1n$, respectively. Strictly speaking, $m = \alpha n$ is not necessarily an integer.

For the asymptotic result of the RTLS estimator, we refer to the conclusion on the LTS estimator developed in [19, Theorem 1] based on Proposition 1.

Assumption 3:

1. The regressors $\{x_k\}$ are deterministic and satisfy $\sum_{k=1}^{n} \|x_k\|^4 = O(n)$, and the matrix $X_n^T X_n/n$ converges to a positive definite matrix Σ as $n \to \infty$.
2. The noise sequence v_i is a sequence of i.i.d. random variables with zero mean and $E(v_i^4) < \infty$. The distribution function $F(\cdot)$ of v_i is absolutely continuous. The density function $f(\cdot)$ of v_i is symmetric, bounded, and positive on $\mathbb{R}$. Moreover, $f(\cdot)$ is strictly decreasing on $[0, +\infty)$ and its first-order and second-order derivatives are bounded.
3. There are distribution functions $H^{(\beta)}(t)$ with $t \in \mathbb{R}$ and $\beta \in \mathbb{R}^p$ such that for all compact set $\mathcal{V} \subset \mathbb{R}^p$,

$$\sup_{\beta\in\mathcal{V}} \sup_{t\in\mathbb{R}} \left| \frac{1}{n} \sum_{k=1}^{n} I\{x_k^T(\beta-\beta^*) \le t\} - H^{(\beta)}(t) \right| = O\big(n^{-\frac{1}{2}}\big),$$

where $I\{\cdot\}$ means the indicator function.

Proposition 2: Consider FIR system (1). Suppose that Assumption 3 holds and $\widehat{\beta}_{n,m}^{rtls}$ is bounded almost surely uniformly over the sample size n. Let $0 < m < \frac{n}{2}$. Then the RTLS estimator $\widehat{\beta}_{n,m}^{rtls} \to \beta^*$ in probability as $n \to \infty$.

IV. Alternating Iterative Numerical Algorithms

This section aims to develop a numerical algorithm that can effectively find a "good" enough candidate for the RTLS estimator based on the well-formulated structure of the optimization problem (9). We emphasize here that finding a good approximation of the OTLS estimator has been a research topic [20] and is not our focus in this paper. To analyze and present the outlier removal ability of the RTLS estimator is the main goal of this work.

Although the optimization problem (9) is nonconvex and its global minimum is generally intractable, two suboptimal problems can be effectively solved by existing optimization algorithms if the variables w and β are treated separately. When w is fixed, the subproblem over the parameter vector β is a least squares problem and has a closed-loop solution. The subproblem for optimizing the weight vector w is a linear programming problem if β is given, which can be effectively solved by several available solvers, e.g., `linprog` in MATLAB. Therefore, here we use the alternating iterative algorithm illustrated in Algorithm 1 to search for a candidate for the RTLS estimator.

Algorithm 1 is actually of the block-coordinated descent type and its convergence to a stationary point is well known.

Proposition 3 ([21]): Consider Algorithm 1 for solving the RTLS problem (9). Then any limit point generated by Algorithm 1 is a stationary point of the optimization problem (9).

Algorithm 1 Alternating Iterative Algorithm for RTLS estimator

Input: The data $\{x_k, y_k\}_{i=1}^n$ and the integer m
Output: β^i and w^i

1: Initialization: Initialize the weight vector $w^0 = (\frac{1}{n}, \cdots, \frac{1}{n})$.
2: Weighted least squares: for a fixed w^{i-1},

$$\beta^i = (X^T W^{i-1} X^T)^{-1} X^T W^{i-1} y$$

with $W^{i-1} \triangleq \mathrm{diag}(w^{i-1})$.
3: Linear programming: for a fixed β^i,

$$w^i = \arg\min_{w\in\mathcal{W}'} J_n(\beta^i, w).$$

4: Set $i = i+1$. Go to step 2 unless some stopping criterion is met.

In practice, the block-coordinated algorithm usually converges to a local minimum, while which local minimum to converge to is unknown as it depends on the initial value of the weight vector w. It is however very likely that the algorithm converges to a global minimum if that global minimum is in a close neighborhood of the uniform distribution $w^\dagger = (\frac{1}{n}, \cdots, \frac{1}{n})$ that is the initial condition of the algorithm. The neighborhood is defined as a ball centered at the uniform distribution with a radius $\mathcal{D}(w) \le \delta$ and moreover this ball contains all the global minimums. In other words, as $\delta \to 0$, the distance between the uniform distribution and any global minimum goes to zero as well. An intuitive interpretation is that when δ or m is small, the two-stage iterative algorithm is likely to converge to a global minimum. Clearly, it is impossible to answer how δ or m is small enough, which depends on the data sets. On the other hand, a very large number of numerical simulations have been carried out and have shown that in every simulation trial, all outliers are successfully identified if their magnitudes are large.

V. Numerical Simulations

We compare the performance of the RTLS and LS estimators against outliers. The system to be simulated is a 5-dimensional FIR system:

$$y_k = \beta_1^* u_{k-1} + \cdots + \beta_5^* u_{k-5} + v_k$$

with an i.i.d. Gaussian input u_k of zero mean and unit variance. Five system parameters $\beta^* = (\beta_1^*, \cdots, \beta_5^*)^T$ of the FIR system are generated uniformly in the interval $[-2, 2]$ independently for each Monte-Carlo run. The data length is $n = 100$ and 100 Monte-Carlo runs are carried out. The noise sequence v_k contains two parts: the normal part is an i.i.d. Gaussian sequence of zero mean and SNR$=$ 20dB, and then outliers are added. The maximum number of outliers is bounded by $m = 0.1n = 10$ in the simulation. The actual number of outliers in each Monte-Carlo run is unknown and generated according to the uniform distribution over

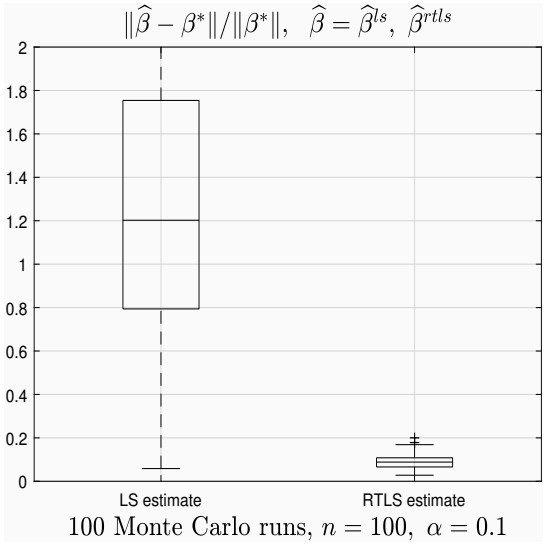

Fig. 1. Parameter estimation error of RTLS and LS, $n = 100,\ \alpha = 0.1$ based on 100 Monte-Carlo runs.

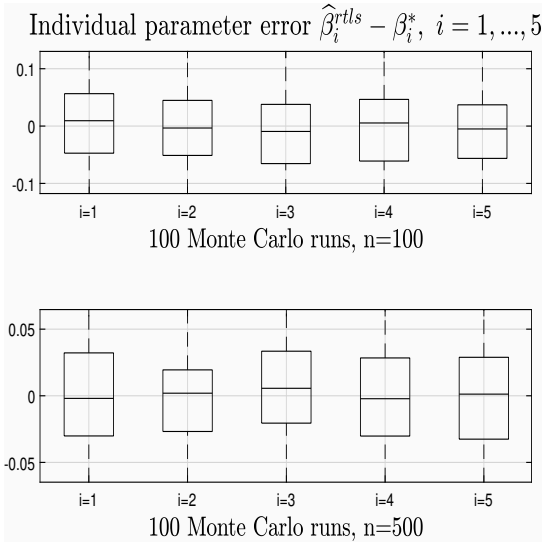

Fig. 2. Individual parameter estimation error $\widehat{\beta}_i^{rtls} - \beta_i^*$, $i = 1, 2, 3, 4, 5$.

$[1, 10]$. The exact locations of outliers are also unknown and again randomly distributed over $\{1, 2, \cdots, n\}$. Finally, if v_k happens to be an outlier, then it is uniformly distributed in the interval $[-100, 100]$. This implies that some outliers could be very large and others are small, but the algorithm does not assume any knowledge on the exact number of outliers and neither their locations nor magnitudes.

Let $\left\|\widehat{\beta}^{ls}-\beta^*\right\|/\left\|\beta^*\right\|$ and $\left\|\widehat{\beta}_{n,m}^{rtls}-\beta^*\right\|/\left\|\beta^*\right\|$ denote the relative parameter estimation errors of the LS and RTLS estimators, respectively. Fig. 1 illustrates the box plots of the parameter estimation errors based on 100 Monte-Carlo runs. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. From Fig. 1, it can be seen that the effect of outliers is largely eliminated by the RTLS estimator while their adverse effect on the LS estimator is significant.

Fig. 2 depicts the box plots of 100 Monte-Carlo runs of the individual error $\widehat{\beta}_{i,m}^{rtls} - \beta_i^*$, $i = 1, 2, 3, 4, 5$ for $n = 100$ (top diagram) and $n = 500$ (bottom diagram), respectively. Further, $\alpha = 0.1$ is adopted for all simulations. Note that the system parameter vector β^* is generated randomly and independently in each Monte-Carlo run and there is no true β^* for all Monte-Carlo runs. So parameter estimation errors represent a better performance indicator.

To see how $\widehat{w}_{n,m}^{rtls}$ is calculated by the RTLS estimator, the result of one simulation run is exhibited in Fig. 3, which shows the generated (unknown) outliers in the top diagram, and the actual outputs in absolute value (dash-dot), corrupted by Gaussian noise and outliers, and the weights $w_{n,m}^{rtls}$ (circle) that are zero derived from the RTLS estimator in the bottom diagram. Note that the maximum number of outliers is $m = 0.1n = 10$ and the actual number of outliers is 9 (unknown). Among them, 8 outliers are significant and one outlier is very small which is indistinguishable with 20dB Gaussian noise. It can be observed that the RTLS estimator identifies

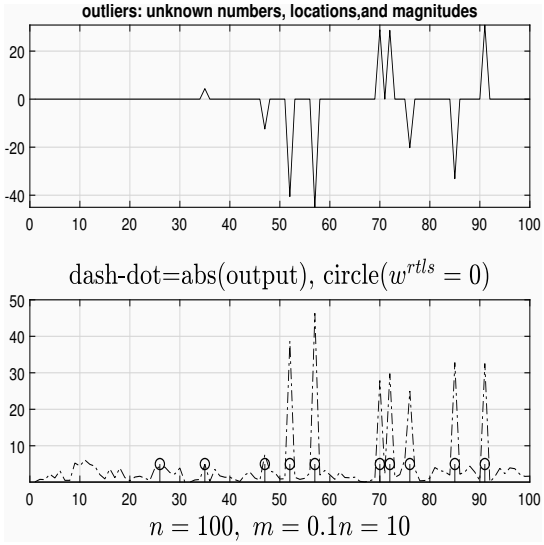

Fig. 3. Actual output and calculated zero weights w^{rtls}.

and removes 10 data points that contains all 9 outliers as expected.

Fig. 4 illustrates the results of the parameter estimation error of the RTLS estimator for $n = 100$, 500, 1000, 2000, respectively, when $\alpha = 0.1$. As expected, as the data length n increases, the parameter estimation error gets smaller.

VI. Conclusion

A relaxed-tilted least squares estimator has been proposed in the paper, aiming to robustify the least squares estimator in the presence of unknown noise contamination. It is a data driven approach that assigns unequal weights to sampled data so that the effect of undesired noise contamination can be mitigated. The proposed estimator tilts the uniform prior on the samples so as to move the uniform distribution in a direction that enjoys the smallest estimation error in the neighborhood of the uniform distribution. (This is the

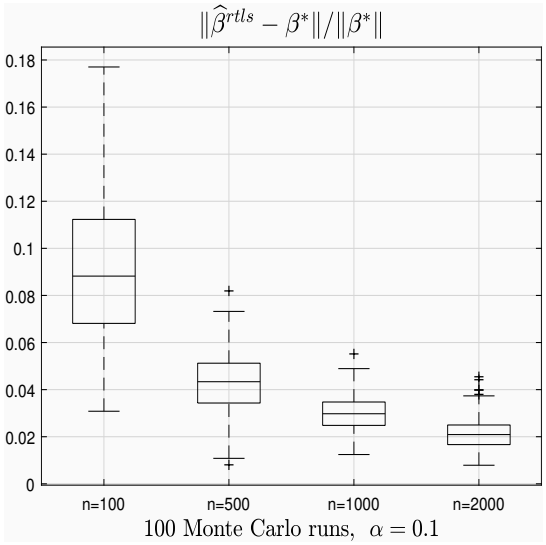

Fig. 4. Parameter estimation error vs n.

property of the tilted least squares estimator). Theoretical analysis including the convergence results of the RTLS estimator has been provided. Extensive numerical simulations have confirmed the theoretical analysis.

REFERENCES

[1] E. W. Bai, H. Cho, R. Tempo, and Y. Ye, "Optimization with few violated constraints for linear bounded error parameter estimation," *IEEE Transactions on Automatic Control*, vol. 47, no. 7, pp. 1067–1077, 2002.

[2] T. Chen and L. Ljung, "Implementation of algorithms for tuning parameters in regularized least squares problems in system identification," *Automatica*, vol. 49, no. 7, pp. 2213–2220, 2013.

[3] B. Mu, E. W. Bai, W. X. Zheng, and Q. Zhu, "A globally consistent nonlinear least squares estimator for identification of nonlinear rational systems," *Automatica*, vol. 77, pp. 322–335, 2017.

[4] M. Verhaegen and V. Verdult, *Filtering and System Identification: A Least Squares Approach*. Cambridge, UK: Cambridge University Press, 2007.

[5] W. X. Zhao and T. Zhou, "Weighted least squares based recursive parametric identification for the submodels of a PWARX system," *Automatica*, vol. 48, no. 6, pp. 1190–1196, 2012.

[6] G. Bottegal, A. Y. Aravkin, H. Hjalmarsson, and G. Pillonetto, "Robust EM kernel-based methods for linear system identification," *Automatica*, vol. 67, pp. 114–126, 2016.

[7] M. Lindfors and T. Chen, "Regularized LTI system identification in the presence of outliers: A variational EM approach," *Automatica*, vol. 121, p. 109152, 2020.

[8] T. Mikosch and C. G. de Vries, "Heavy tails of OLS," *Journal of Econometrics*, vol. 172, no. 2, pp. 205–221, 2013.

[9] J. Erickson, S. Har-Peled, and D. M. Mount, "On the least median square problem," *Discrete & Computational Geometry*, vol. 36, pp. 593–607, 2006.

[10] P. J. Rousseeuw, "Least median of squares regression," *Journal of the American Statistical Association*, vol. 79, no. 388, pp. 871–880, 1984.

[11] S. P. Ellis, "Instability of least squares, least absolute deviation and least median of squares linear regression," *Statistical Science*, vol. 13, no. 4, pp. 337–350, 1998.

[12] H. L. Harter, "Nonuniqueness of least absolute values regression," *Communications in Statistics - Theory and Methods*, vol. 6, no. 9, pp. 829–838, 1977.

[13] W. Xu, E. W. Bai, and M. Cho, "System identification in the presence of outliers and random noises: A compressed sensing approach," *Automatica*, vol. 50, no. 11, pp. 2905–2911, 2014.

[14] J. Y. Audibert and O. Catoni, "Robust linear least squares regression," *The Annals of Statistics*, vol. 39, no. 5, pp. 2766–2794, 2011.

[15] E. Choi, P. Hall, and B. Presnell, "Rendering parametric procedures more robust by empirically tilting the model," *Biometrika*, vol. 87, no. 2, pp. 453–465, 2000.

[16] J. Duchi and H. Namkoong, "Variance-based regularization with convex objectives," *Journal of Machine Learning Research*, vol. 19, pp. 1–55, 2018.

[17] S. P. Ellis and S. Morgenthaler, "Leverage and breakdown in l_1 regression," *Journal of the American Statistical Association*, vol. 87, no. 417, pp. 143–148, 1992.

[18] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[19] J. A. Víšek, "The least trimmed squares Part I: Consistency," *Kybernetika*, vol. 42, pp. 1–36, 2006.

[20] J. Agulló, "New algorithms for computing the least trimmed squares regression estimator," *Computational Statistics & Data Analysis*, vol. 36, no. 4, pp. 425–439, 2001.

[21] L. Grippo and M. Sciandrone, "On the convergence of the block nonlinear Gauss-Seidel method under convex constraints," *Operations Research Letters*, vol. 26, no. 3, pp. 127–136, 2000.