Mozafari-Majd, Emadaldin; Koivunen, Visa

Robust variable selection and distributed inference using t-based estimators for large-scale data

Published in:
28th European Signal Processing Conference, EUSIPCO 2020 - Proceedings

DOI:
10.23919/Eusipco47968.2020.9287773

Published: 01/01/2020

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Mozafari-Majd, E., & Koivunen, V. (2020). Robust variable selection and distributed inference using t-based estimators for large-scale data. In *28th European Signal Processing Conference, EUSIPCO 2020 - Proceedings* (pp. 2453-2457). Article 9287773 (European Signal Processing Conference). European Association For Signal and Image Processing. https://doi.org/10.23919/Eusipco47968.2020.9287773

Robust variable selection and distributed inference using τ-based estimators for large-scale data

Emadaldin Mozafari-Majd and Visa Koivunen
Aalto University, Espoo, Finland
emadaldin.mozafarimajd@aalto.fi, visa.koivunen@aalto.fi

***Abstract*—In this paper, we address the problem of performing robust statistical inference for large-scale data sets whose volume and dimensionality maybe so high that distributed storage and processing is required. Here, the large-scale data are assumed to be contaminated by outliers and exhibit sparseness. We propose a distributed and robust two-stage statistical inference method. In the first stage, robust variable selection is done by exploiting τ-Lasso to find the sparse basis in each node with distinct subset of data. The selected variables are communicated to a fusion center (FC) in which the variables for the complete data are chosen using a majority voting rule. In the second stage, confidence intervals and parameter estimates are found in each node using robust τ-estimator combined with bootstrapping and then combined in FC. The simulation results demonstrate the validity and reliability of the algorithm in variable selection and constructing confidence intervals even if the estimation problem in the subsets is slightly underdetermined.**

***Index Terms*—statistical inference, robust, sparse, high-dimensional, large-scale data, variable selection, bootstrap**

I. Introduction

The proliferation of heterogeneous large-scale data sets generated in an unprecedented volume and speed have led to dramatic changes in data storage, processing and inferential methods. In particular, distributed solutions with multicore and cloud computing platforms facilitate the storage and processing of high volume of data. Conventional statistical inference and learning methods can not accommodate processing of such large-scale data sets due to a lack of scalability. In order to address the scalability, statistical methods are needed to be compatible with distributed computing and storage platforms. Moreover, in many practical applications often large-scale data sets are contaminated by outliers, and bias in estimated parameters and confidence intervals may be significant. In order to achieve a desired accuracy and reliability, robust statistical methods are preferred, see [1]. A statistically robust, fast and scalable bootstrap compatible with distributed architectures was introduced in [2]. The inference method employs one-step estimators in combination with fixed-point equations to faithfully estimate parameters of interest and measures of uncertainty in terms of confidence intervals.

In many large-scale and high-dimensional data analysis problems, the number of explaining variables may be of the same order or larger than the number of observations. Hence, the problem may become ill-posed and some regularization may be necessary. High-dimensional data sets are arising in many applications such as genomics, finance, face recognition as well as medicine. To obtain valid inference in high-dimensional scenarios, it is required to impose structural constraints such as sparsity and low-rank on the large-scale data.

In this paper, we proposed a robust and scalable two-stage statistical inference method concerning with high-dimensional data and compatible with distributed architecture. In order to achieve scalability, data are stored and processed locally at each node. This may be achieved by subdividing the full data into smaller distinct subsets, for example by resampling without replacement. In the first stage of the proposed method, robust variable selection is done by employing τ-Lasso [3], [4] to identify the sparse basis for each distinct subset of data. The selected variables are then communicated to a cloud or fusion center that selects the variables for all the nodes via majority voting rule. The selected basis is then communicated back to each distributed node. The estimates of parameters and confidence intervals at each node are found by using a robust extension of the Bag of Little Bootstraps (BLB) method [5] proposed in this paper. The bootstrap replicates are computed using a robust τ-estimator. The estimated confidence intervals from all nodes are combined in the fusion center by using trimmed means. Tuning-free estimation of bootstrap replicates and exploiting multinomial weighting in τ-estimation offers high scalability. The simulations demonstrate the high reliability of the proposed algorithm both in variable selection as well as finding point estimates and confidence intervals. The method has comparable performance to MM-estimator based inference in mildly under-determined scenarios (subsets of data). Moreover, it outperforms the MM-Lasso based method in variable selection performance, having lower false positive rate under given scenario.

This paper is organized as follows: In section 2, the basics of statistical inference for large-scale data and brief explanation of the proposed method is introduced. In section 3 and 4, the robust variable selection and distributed inference is explained in details. Section 5 provides simulation studies and investigates the performance of the proposed method.

II. Preliminaries and Prior Works

In this paper, the large-scale data under investigation follow a linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{v}$ where $\mathbf{X} \in \mathbb{R}^{n \times p}$

This work has been supported by the Academy of Finland, grant "Statistical Signal Processing Theory and Computational Methods for Large Scale Data Analysis".

denotes a regression matrix, $\mathbf{y} \in \mathbb{R}^n$ is a response vector, $\mathbf{v} \in \mathbb{R}^n$ is a measurement noise and $\boldsymbol{\beta} \in \mathbb{R}^p$ denotes a sparse parameter vector with $k_s = |\mathcal{S}|$ non-zero entries and $\mathcal{S} = \{j \in 1, \cdots, p : \beta_j \neq 0\}$. Because of the large volume, the data $\mathbf{Y} = \big(\mathbf{y}, \mathbf{X}\big)$ are divided into s smaller distinct subsets of data $\check{\mathbf{Y}}^{(i)} = \big(\check{\mathbf{y}}^{(i)}, \check{\mathbf{X}}^{(i)}\big) \in \mathbb{R}^{b \times (p+1)}, i = 1, \cdots, s$ that can be stored and processed separately. The subsets are formed by resampling without replacement from the rows of the complete data set where $b = \{\lfloor n^\gamma \rfloor | \gamma \in [0.5, 1)\}$. The same situation would occur if subsets of data are stored on s storage and computing nodes and each node contains b observations. Sometimes, the value of b may be such that $p \approx b$, $p/b > 1$ or data may be contaminated by outliers. Statistically robust regularized estimators can be used to address these problems simultaneously [6], [7] and [8]. In high-dimensional statistics, performing inference tasks such as finding confidence intervals or hypothesis testing require further effort to reduce the bias introduced by regularization. State-of-art statistical inference methods in high-dimensional statistics follow two avenues, debiased-Lasso [9] and post-Lasso estimators [10] and [11]. The debiased-Lasso compensates for the bias introduced by the Lasso estimator and provides valid solutions to finding confidence intervals and hypothesis testing. Alternatively, post-Lasso estimators offer inference solutions in two stages where in the first stage a subset of variables is selected and in the second stage bootstrapping least square type estimators is employed to perform the actual inference.

In order to deal with potentially under-determined models and perform robust variable selection in the presence of outliers, we proposed a statistical inference technique in [12], inspired by Bolasso [13] and post-Lasso estimators. In this paper, we propose a two-stage distributed statistical inference method where robust variable selection is performed in the first stage. The actual inference is done in the second stage using statistically robust bootstrapping and the variables selected in the first stage. The robust variable selection method exploits the τ-Lasso method [3], [4] to find the sparse basis for each of the s distinct data sets of b observations. The selection results from each node are fused in a cloud or fusion center by using a k-out-of-p voting scheme to choose the variables for the complete large data set. The chosen basis is communicated back to each distributed computing and storage node and used in the second stage of the inference. A statistically robust extension of Bag of Little Bootstraps (BLB) [11] is employed to find parameter estimates and confidence intervals for the selected basis in each node. Bootstrap replicates are computed using a robust low-dimensional τ-estimator of regression [14]. The bootstrap percentile method is used to estimate confidence intervals associated with the selected variables. The estimated confidence intervals from each node are communicated to the cloud or fusion center for the inference on complete large-scale data. The confidence intervals are combined in the fusion center by applying trimmed mean over the lower and upper bound of confidence intervals. The details of the proposed method is presented in the following sections.

III. DISTRIBUTED ROBUST VARIABLE SELECTION

In this section, the proposed consistent support recovery algorithm employed in the first stage of the inference method is described in detail. The main steps of the algorithm are summarized in Figure 1.

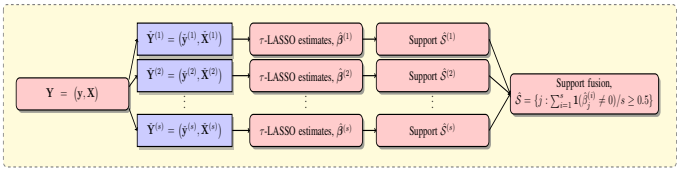

Fig. 1: First Stage: The support recovery algorithm extracts the support from different bags using sparsity and finds the support of the parameter by using the majority rule.

In the first step, the large-scale data are assumed to be divided into s distinct subsets by resampling without replacement. The robust τ-Lasso estimator [3], [4] is applied to each distinct subset of observations $\check{\mathbf{Y}}^{(i)} = (\check{\mathbf{y}}^{(i)}, \check{\mathbf{X}}^{(i)}), i = 1, ..., s$ as follows:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(i)} &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, \mathcal{L}(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \Big([\hat{\sigma}_\tau^{(i)}]^2 + \lambda \|\boldsymbol{\beta}\|_{\ell_1} \Big) \\ &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \Big(\frac{[\hat{\sigma}_b^{(i)}]^2}{b} \sum_{l=1}^{b} \rho_1 \Big(\frac{\check{r}_l^{(i)}(\boldsymbol{\beta})}{\hat{\sigma}_b^{(i)}} \Big) + \lambda \|\boldsymbol{\beta}\|_{\ell_1} \Big),\end{aligned} \tag{1}$$

where λ determines the level of sparsity imposed by the ℓ_1-norm penalty term, $\check{\mathbf{r}}^{(i)}(\boldsymbol{\beta}) = \check{\mathbf{y}}^{(i)} - \check{\mathbf{X}}^{(i)}\boldsymbol{\beta}$, $\rho_1(\cdot)$ is an even and bounded function satisfying the properties of bounded ρ-function defined by Maronna et al. [15]. In this work, Tukey's bisquare ρ-function is used, defined as $\rho(t) = 1 - \big(1 - (t/c)^2\big)^3 \mathbf{1}(|t| \leq c)$ where c is a tuning constant that trades off between robustness and efficiency. $\hat{\sigma}_\tau^{(i)} = \hat{\sigma}_\tau(\check{\mathbf{r}}^{(i)}(\boldsymbol{\beta}))$ and $\hat{\sigma}_b^{(i)} = \hat{\sigma}_b(\check{\mathbf{r}}^{(i)}(\boldsymbol{\beta}))$ are the τ-scale and M-scale of the residual vector $\check{\mathbf{r}}^{(i)}(\boldsymbol{\beta})$, respectively. Here, the M-scale estimate of the residual vector satisfies $(1/b)\sum_{l=1}^{b} \rho_0(\check{r}_l^{(i)}(\boldsymbol{\beta})/\hat{\sigma}_b^{(i)}) = \delta_1$ where $\rho_0(\cdot)$ is similarly an even and bounded ρ-function and δ_1 controls the break-down point of the estimator. The constants c_0 and c_1 ($c_1 \geq c_0$) are tuned according to $\mathbb{E}[\rho_0(t)] = \delta^*$ and $(\mathbb{E}[\psi_\tau'(t)])^2/\mathbb{E}[\psi_\tau^2(t)] = \zeta^*$ to ensure the desired high break-down point δ^* and efficiency ζ^* are satisfied for linear models with normal errors, simultaneously. It is assumed that $t \sim \mathcal{N}(0,1)$, and $\psi_\tau(t) = \bar{w}_\tau \psi_0(t) + \psi_1(t)$, $\psi_0(t) = \partial\rho_0(t)/\partial t$ and $\psi_1(t) = \partial\rho_1(t)/\partial t$, respectively. $\bar{w}_\tau$ is given by $\bar{w}_\tau = \big(2\mathbb{E}[\rho_1(t)] - \mathbb{E}[\psi_1(t)t]\big)/\mathbb{E}[\psi_0(t)t]$. The objective function given in equation [1] consists of a non-convex term and a non-smooth ℓ_1-norm penalty term. Therefore, minimizing such an objective function is not a trivial task and we address this issue by taking the generalized gradient of the objective function, defined as $\partial_{\boldsymbol{\beta}}([\hat{\sigma}_\tau^{(i)}]^2 + \lambda\|\boldsymbol{\beta}\|_{\ell_1})$. Here, the generalized gradient of the non-convex smooth, continuously differentiable term $\partial_{\boldsymbol{\beta}}[\hat{\sigma}_\tau^{(i)}]^2$ is identical to its gradient $\nabla_{\boldsymbol{\beta}}[\hat{\sigma}_\tau^{(i)}]^2$ and the generalized gradient of the non-smooth,

convex term $\partial_{\boldsymbol{\beta}}(\lambda\|\boldsymbol{\beta}\|_{\ell_1})$ coincides with its subdifferential [7], [16]. As the objective function is locally Lipschitz, any point $\boldsymbol{\beta}_0$ at which $\mathbf{0} \in \partial_{\boldsymbol{\beta}}([\hat{\sigma}_\tau^{(i)}]^2 + \lambda\|\boldsymbol{\beta}\|_{\ell_1})$ is a local minimum of the τ-Lasso estimation problem. To find local minima of the given estimation problem, the generalized gradient of the objective function is taken wrt $\boldsymbol{\beta}$. It is given by

$$\partial_{\boldsymbol{\beta}}\mathcal{L}(\boldsymbol{\beta}) = -\frac{\hat{\sigma}_b^{(i)}}{b}\sum_{l=1}^{b} w_l^{(i)}\check{r}_l^{(i)}(\boldsymbol{\beta})\check{\mathbf{x}}_{[l]}^{(i)} + \lambda\partial_{\boldsymbol{\beta}}(\|\boldsymbol{\beta}\|_{\ell_1}), \quad (2)$$

where the generalized gradient of the objective function turns out to be equivalent to the sub-gradient of the weighted least square penalized by ℓ_1-norm except that the weights $w_l^{(i)}(\boldsymbol{\beta})$ here depend on the unknown $\boldsymbol{\beta}$. Therefore, we can reformulate the optimization problem as follows:

$$\hat{\boldsymbol{\beta}}^{(i)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\boldsymbol{\Omega}^{(i)}(\check{\mathbf{y}}^{(i)} - \check{\mathbf{X}}^{(i)}\boldsymbol{\beta})\|_{\ell_2}^2 + \lambda^{'}\|\boldsymbol{\beta}\|_{\ell_1}, \quad (3)$$

where $\lambda^{'} = 2b\lambda/\hat{\sigma}_b^{(i)}$, $\boldsymbol{\Omega}^{(i)}$ is a diagonal matrix whose entries on diagonal are $\sqrt{w_l^{(i)}}$ and $w_l^{(i)}$ is given by,

$$\begin{aligned} w_l^{(i)} &= \frac{\left[w_\tau^{(i)}\psi_0(\tilde{r}_l^{(i)}) + \psi_1(\tilde{r}_l^{(i)})\right]}{\check{r}_l^{(i)}}, \\ w_\tau^{(i)} &= \frac{\sum_{l=1}^{b}\left[2\rho_1(\tilde{r}_l^{(i)}) - \psi_1(\tilde{r}_l^{(i)})\tilde{r}_l^{(i)}\right]}{\sum_{l=1}^{b}\psi_0(\tilde{r}_l^{(i)})\tilde{r}_l^{(i)}}. \end{aligned} \quad (4)$$

Here, the notation $\check{r}_l^{(i)}$ is a shorthand for $\check{r}_l^{(i)}(\boldsymbol{\beta})$ and $\tilde{r}_l^{(i)} = \check{r}_l^{(i)}/\hat{\sigma}_b^{(i)}$. The optimization problem defined in equation (3) is solved by using the iteratively reweighted Lasso (IR-WLASSO) that alternates between finding the weight matrices $\boldsymbol{\Omega}^{(i)}$, refining $\hat{\sigma}_b^{(i)}$ and updating $\hat{\boldsymbol{\beta}}^{(i)}$. Each node computes an estimate of the parameter vector $\hat{\boldsymbol{\beta}}^{(i)}$ and communicates its selected support $\hat{\mathcal{S}}^{(i)} = \{j \in 1, \cdots, p : \hat{\beta}_j^{(i)} \neq 0\}$ to the cloud or fusion center which uses the majority voting rule to select the variables for the complete large-scale data set. The chosen basis $\hat{\mathcal{S}}$ is communicated back to each node and used in the second stage of inference. $\hat{\mathcal{S}} = \{j : \sum_{i=1}^{s}\mathbf{1}(\hat{\beta}_j^{(i)} \neq 0)/s \geq 0.5\}$, i.e. if a parameter is in the support within the majority of subsets, it is selected to the support for the complete data set. In order to tune the regularization parameter λ, we create a grid of N_{λ_τ} lambdas for tuning the regularization parameter of τ-Lasso and N_{λ_S} lambdas for tuning the regularization parameter of S-Lasso and find the optimal values using the method explained in our earlier paper [12].

IV. INFERENCE USING τ-ESTIMATOR AND BLB

In this section, we describe the inference performed in stage 2 of the proposed approach. We develop a new statistically robust extension of Bag of Little Bootstraps (BLB) [5] to address the robustness and avoiding bias in estimation of $\boldsymbol{\beta}$ and confidence intervals. The algorithm operates by using the explaining variables selected in stage 1. Hence, the columns of $\check{\mathbf{X}}^{(i)}$ that correspond to variables excluded from $\hat{\mathcal{S}}$ are discarded. The algorithm generates B bootstrap samples for each subset of data $\check{\mathbf{Y}}_{\hat{\mathcal{S}}}^{(i)} = (\check{\mathbf{y}}^{(i)}, \check{\mathbf{X}}_{\hat{\mathcal{S}}}^{(i)}), i = 1, \cdots, s$ where the multiplicity of observations in the bootstrap sample is determined by a random weight vector $\boldsymbol{\omega}^{*(ij)} \in \mathbb{R}^b, j = 1, \cdots, B$ drawn from a multinomial distribution $\big(n, (1/b)\mathbf{1}_b\big)$. Here, the bootstrap replicates $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}^{*(ij)}$ are estimated using a robust τ-estimator of regression [14] as follows:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}^{*(ij)} &= \underset{\boldsymbol{\beta}_{\hat{\mathcal{S}}}^{*}}{\operatorname{argmin}} \Big(\frac{[\hat{\sigma}_b^{*(ij)}]^2}{n}\sum_{l=1}^{b}\omega_l^{*(ij)}\rho_1(\tilde{r}_l^{*(ij)})\Big), \\ &= \underset{\boldsymbol{\beta}_{\hat{\mathcal{S}}}^{*}}{\operatorname{argmin}} \|\boldsymbol{\Omega}^{*(ij)}(\check{\mathbf{y}}^{(i)} - \check{\mathbf{X}}_{\hat{\mathcal{S}}}^{(i)}\boldsymbol{\beta}_{\hat{\mathcal{S}}}^{*})\|_{\ell_2}^2, \end{aligned} \quad (5)$$

where M-scale of residuals, $\hat{\sigma}_b^{*(ij)}$, satisfies $(1/n)\sum_{l=1}^{b}\omega_l^{*(ij)}\rho_0(\tilde{r}_l^{*(ij)}) = \delta_2$, $\tilde{r}_l^{*(ij)} = \check{r}_l^{(i)}(\boldsymbol{\beta}_{\hat{\mathcal{S}}}^{*})/\hat{\sigma}_b^{*(ij)}$ and $\boldsymbol{\Omega}^{*(ij)} = \text{diag}(\sqrt{\mathbf{w}^{*(ij)} \odot \boldsymbol{\omega}^{*(ij)}})$. $w_l^{*(ij)}$ is given by

$$\begin{aligned} w_\tau^{*(ij)} &= \frac{\sum_{l=1}^{b}\omega_l^{*(ij)}\Big[2\rho_1(\tilde{r}_l^{*(ij)}) - \psi_1(\tilde{r}_l^{*(ij)})\tilde{r}_l^{*(ij)}\Big]}{\sum_{l=1}^{b}\omega_l^{*(ij)}\psi_0(\tilde{r}_l^{*(ij)})\tilde{r}_l^{*(ij)}}, \\ w_l^{*(ij)} &= \frac{\left[w_\tau^{*(ij)}\psi_0(\tilde{r}_l^{*(ij)}) + \psi_1(\tilde{r}_l^{*(ij)})\right]}{\check{r}_l^{(i)}(\boldsymbol{\beta}_{\hat{\mathcal{S}}}^{*})}. \end{aligned} \quad (6)$$

The iteratively reweighted least square (IR-WLS) is employed to deal with dependence of weights on the unknown $\boldsymbol{\beta}_{\hat{\mathcal{S}}}^{*}$. Once $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}^{*(ij)}$s are robustly estimated for each bootstrap replicate, the confidence intervals can be estimated using the bootstrap percentile method, for example. The estimated confidence intervals are transmitted from each computing node to the fusion center to perform the inference for the complete large-scale data. In the fusion center, the confidence intervals for the complete large-scale data are estimated by applying μ-trimmed mean over upper bounds and lower bounds of transmitted confidence intervals as follows:

$$\overline{\mathbf{CI}}_\mu = \frac{1}{s-2\kappa}\sum_{i=\kappa+1}^{s-\kappa}[\mathbf{CI}_{\hat{\mathcal{S}}}]^{\star(i)} \quad (7)$$

where $\mu \in [0, 1/2)$ denotes the proportion of entries to be discarded from lower bound and upper bound, $\kappa = [s\mu]$ determines the number of entries within upper bound and lower bound each to be truncated ([.] stands for the integer part) and $[\mathbf{CI}_{\hat{\mathcal{S}}}]^{\star(i)}$ denotes the order statistics. The overall procedure to construct confidence intervals is summarized in the following steps:

1) Generate B bootstrap samples for each distinct subset of data, $\check{\mathbf{Y}}_{\hat{\mathcal{S}}}^{*(ij)} = \big(\check{\mathbf{y}}^{(i)}, \check{\mathbf{X}}_{\hat{\mathcal{S}}}^{(i)}; \boldsymbol{\omega}^{*(ij)}\big), j = 1, ..., B$
2) Compute bootstrap τ estimates, $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}^{*(ij)}$
3) Compute $(1-\alpha)\%$ confidence intervals $\mathbf{CI}_{\hat{\mathcal{S}}}^{*(i)}$ for each subset of data using the bootstrap percentile method,
4) Fuse the lower and upper confidence bounds by using μ-trimmed mean, $\overline{\mathbf{CI}}_\mu = \frac{1}{s-2\kappa}\sum_{i=\kappa+1}^{s-\kappa}[\mathbf{CI}_{\hat{\mathcal{S}}}]^{*(i)}$

V. RESULTS

In this section, the performance of the proposed method is investigated in simulations considering both variable selection

and inference. In particular, identifying sparse basis correctly, statistical robustness, the quality of the parameter estimates and confidence intervals are studied. Different proportions of outliers and a variety of SNR levels are used in simulations. The performance is compared to an inference method using MM-Lasso and MM-estimators in [12]. It is assumed the large-scale data follow a linear regression model where the parameter vector $\boldsymbol{\beta} \in \mathbb{R}^p$ is sparse with $k_s = 40$ non-zero entries. $\boldsymbol{\beta}_{\mathcal{S}}$ is set to $3 \times \mathbf{1}_{\mathcal{S}}$ and their positions are chosen randomly. The regression matrix is a Toeplitz matrix with i.i.d rows drawn from a multivariate Gaussian $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\Sigma_{ij} = \rho^{|i-j|}$ where ρ is a constant. The measurement noise vector, $\mathbf{v}$, is an additive white Gaussian noise with a variance (AWGN) $\sigma_v^2 = \|\mathbf{X}\boldsymbol{\beta}\|_{\ell_2}^2 10^{-\text{SNR}/10}/n$ (SNR in dB). The outliers are introduced by randomly choosing the observations in $\mathbf{y}$ and replacing them with random values chosen from a standard Gaussian distribution with $\sigma_e = 250$. We set the simulation parameters as follows: $n = 8100$, $p = 270$, $b = 225$, $\gamma = 0.602$, $c_1 = 6.04$ for τ-Lasso and $c_1 = 3.44$ for MM-Lasso, $s = 36$, $N_{\lambda_{\text{S}}} = 50$ (number of points in the grid of lambda for S), $N_{\lambda_\tau} = 70$ (number of points in the grid of lambda for τ), $N_\tau = 7$ (maximum number of iterations), $B = 100$, $\rho = 0.5$, $\mu = \mathcal{OP}/2$, $\kappa = 1/2(\mathcal{OP} \times s)$ and $\alpha = 0.1$. The tuning constant c_1 is adjusted to provide 85% efficiency under Gaussian errors for bootstrap τ-estimates and bootstrap MM-estimates ($c_1 = 3.44$ for MM). The remaining of parameter set-up for MM-Lasso and MM estimator will be identical to the above set-up.

Across all simulations, it is assumed the linear regression model has an intercept component and all columns of the augmented regression matrix $[\mathbf{1}_{b\times 1}\check{\mathbf{X}}^{(i)}]$ are robustly standardized by centring the columns using a bisquare location estimator and scaling them using bisquare scale estimators [17]. The response vector $\check{\mathbf{y}}^{(i)}$ is centred using the bisquare location estimator [17]. In order to tune c_0 for the initial S-Lasso estimator and M-scale of residuals, we set $\delta_1 = 0.25 \times \mathbf{1}(\mathcal{OP} < 0.2) + (\mathcal{OP} + 0.05) \times \mathbf{1}(\mathcal{OP} \geq 0.2)$, which controls the breakdown point according to Theorem 4.1 in [7]. It is well justified to set δ_1 to a higher value than the outlier proportion $\mathcal{OP}$ and keep a safety gap large enough between δ_1 and $\mathcal{OP}$. This would mitigate the chances of breaking the estimator when the subsets of data contain higher proportion of outliers than the nominal $\mathcal{OP}$. On the other hand, there is a trade-off between robustness and bias in M-scale estimators which is more pronounced in the presence of sparsity [12]. When tuning c_0 for the initial S estimator and M-scale of residuals for bootstrap samples, we need to increase the safety gap even more $\delta_2 = \mathcal{OP} + 0.1$. Because bootstrapping may lead to having larger proportion of outliers in replicate data sets than δ_2. The proposed algorithm was implemented in MATLAB except the estimation of initial S-Lasso which was done in R using PENSE [7], (https://cran.r-project.org/package=pense). In addition, we used the Dual Augmented Lagrangian [18] implementation to solve the IR-WLASSO.

We run 10 trials of the above simulation setup for which a random realization of outliers is used at each trial. $\mathbf{X}$, $\boldsymbol{\beta}$ and $\mathbf{v}$ are kept fixed across all trials. The simulation results of variable selection performance are reported by taking average of 10 trials and then shown in Table I and II. As it can be observed, the proposed algorithm exhibits a very reliable performance in variable selection and the true positive rate is 100%. It's worth noting that even though the number of false positives is remarkably low, they can be further reduced in the second stage by rejecting variables when corresponding CI contains zero. In addition, the proposed robust variable selection outperforms the robust variable selection employing MM-Lasso estimator under the given scenario.

The reliability of parameter and CI estimates is demonstrated in Figure 2 and Table III. The performance is compared to inference method using Conventional Lasso and least square estimators. In simulations for outlier free scenario, 10 random realizations of measurement noise is generated and $\mathbf{X}$ and $\boldsymbol{\beta}$ are kept fixed across all trials. The estimated confidence intervals are obtained by averaging the CIs over lower and upper bounds from 10 trials at SNR level of 20 dB. Employing non-robust estimators results in a false positive rate of 100%, indicating that all zero entries were incorrectly identified as non-zero and the estimated CIs are extremely wide and hence non-informative. RMSE of the parameter estimates, as given in Figure 2, are significantly large for this case. As it can be observed, the proposed method provides robust estimates of CIs that are just slightly inflated when the proportion of outliers increases. This can be confirmed by RMSE values of the parameter estimates given in Table III and how well they are concentrated about the true values. Only in very few cases the true parameter values were not within the estimated CI. Comparing the proposed algorithm to the inference method employing MM-Lasso and MM estimators indicates that with the proposed method, confidence intervals are slightly larger and provide better coverage properties under given scenario for higher proportion of outliers.

TABLE I: Contingency table for under-determined case ($p/b = 1.2$) and $\mathcal{OP} = 0.3$, $\text{CER}_{\text{MM}} = 0.0015 > \text{CER}_\tau = 0.0011$, ($\text{RER}_{\text{MM}} = 0.9985 < \text{RER}_\tau = 0.9989$), SNR=30 dB.

True \ Classified	τ	Sparse	Zero
	Sparse	100	0
	Zero	0.13	99.87

True \ Classified	MM	Sparse	Zero
	Sparse	100	0
	Zero	0.17	99.83

TABLE II: Contingency table for under-determined case ($p/b = 1.2$) and $\mathcal{OP} = 0.3$, $\text{CER}_{\text{MM}} = 0.0059 > \text{CER}_\tau = 0.0015$, ($\text{RER}_{\text{MM}} = 0.9941 < \text{RER}_\tau = 0.9985$), SNR=20 dB.

True \ Classified	τ	Sparse	Zero
	Sparse	100	0
	Zero	0.17	99.83

True \ Classified	MM	Sparse	Zero
	Sparse	100	0
	Zero	0.70	99.30

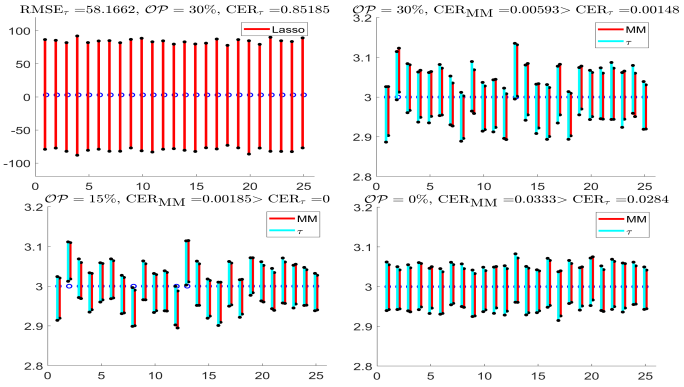

Fig. 2: The wide confidence intervals obtained by using non-robust estimators convey no information whereas the confidence intervals obtained by using robust estimators are minimally affected when the proportion of outlier increases. Reliable parameter estimates and their confidence intervals are obtained even in the presence of outliers ($\beta_{\mathcal{S}}$ blue dots). Moreover, the proposed method achieves lower CER values compared to the inference method using MM-based estimators.

TABLE III: RMSE values of the parameter estimates for different proportion of outliers. The proposed method using τ-based estimators have almost equal performance in comparison to the inference method using MM-based estimators in terms of RMSE values with a slightly lower RMSE in higher outlier ratio and in contrast a slightly higher RMSE in lower outlier ratios.

RMSE	MM	τ
$\mathcal{OP} = 0\%$	0.157	0.163
$\mathcal{OP} = 15\%$	0.153	0.155
$\mathcal{OP} = 30\%$	0.187	0.186

VI. CONCLUSION

In this paper, a robust and distributed two-stage statistical inference method for large-scale data sets having sparse underlying structure and outlying observations was proposed. The performance of the algorithm was studied in terms of variable selection, measured by Classification Error Rate (CER), Recovery Error Rate (RER) and contingency table and accuracy of confidence intervals, quantified by RMSE of parameter estimates, coverage and box-plots. In addition, the proposed algorithm was compared to the non-robust counterpart and a robust inference method based on MM-estimators. The results show the high reliability of the proposed algorithm in variable selection, computing parameter estimates and finding confidence intervals while avoiding the bias introduced by regularization. The proposed method outperformed the inference method using MM-Lasso estimator in terms of variable selection, having lower false positive rate under the given scenario in simulations. The RMSE performances of these two methods are almost equal.

REFERENCES

[1] Abdelhak M Zoubir, Visa Koivunen, Esa Ollila, and Michael Muma, *Robust statistics for signal processing*, Cambridge University Press, 2018.

[2] Shahab Basiri, Esa Ollila, and Visa Koivunen, "Robust, scalable, and fast bootstrap method for analyzing large scale data," *IEEE Transactions on Signal Processing*, vol. 64, no. 4, pp. 1007–1017, 2015.

[3] Marta Martinez-Camara, Michael Muma, Abdelhak M Zoubir, and Martin Vetterli, "A new robust and efficient estimator for ill-conditioned linear inverse problems with outliers," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 3422–3426.

[4] Marta Martinez-Camara, Michael Muma, Benjamin Bejar, Abdelhak M Zoubir, and Martin Vetterli, "The regularized tau estimator: A robust and efficient solution to ill-posed linear inverse problems," *arXiv preprint arXiv:1606.00812*, 2016.

[5] Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael I Jordan, "A scalable bootstrap for massive data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 4, pp. 795–816, 2014.

[6] Ezequiel Smucler and Victor J Yohai, "Robust and sparse estimators for linear regression models," *Computational Statistics & Data Analysis*, vol. 111, pp. 116–130, 2017.

[7] Gabriela V Cohen Freue, David Kepplinger, Matías Salibián-Barrera, and Ezequiel Smucler, "Pense: A penalized elastic net s-estimator," 2017.

[8] Nam H Nguyen and Trac D Tran, "Robust lasso with missing and grossly corrupted observations," *IEEE transactions on information theory*, vol. 59, no. 4, pp. 2036–2058, 2012.

[9] Adel Javanmard and Andrea Montanari, "Confidence intervals and hypothesis testing for high-dimensional regression," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2869–2909, 2014.

[10] Hanzhong Liu, Bin Yu, et al., "Asymptotic properties of lasso+ mls and lasso+ ridge in sparse high-dimensional linear regression," *Electronic Journal of Statistics*, vol. 7, pp. 3124–3169, 2013.

[11] Hanzhong Liu, Xin Xu, and Jingyi Jessica Li, "A bootstrap lasso+ partial ridge method to construct confidence intervals for parameters in high-dimensional sparse linear models," *arXiv preprint arXiv:1706.02150*, 2017.

[12] Emadaldin Mozafari Majd and Visa Koivunen, "Robust, sparse and scalable inference using bootstrap and variable selection fusion," in *2019 8th IEEE International Workshop on Compuational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, 2019.

[13] Francis R Bach, "Bolasso: model consistent lasso estimation through the bootstrap," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 33–40.

[14] Victor J Yohai and Ruben H Zamar, "High breakdown-point estimates of regression by means of the minimization of an efficient scale," *Journal of the American statistical association*, vol. 83, no. 402, pp. 406–413, 1988.

[15] Ricardo A Maronna, R Douglas Martin, Victor J Yohai, and Matías Salibián-Barrera, *Robust statistics: theory and methods (with R)*, John Wiley & Sons, 2019.

[16] Frank H Clarke, *Optimization and nonsmooth analysis*, vol. 5, Siam, 1990.

[17] Ricardo A Maronna, "Robust ridge regression for high-dimensional data," *Technometrics*, vol. 53, no. 1, pp. 44–53, 2011.

[18] Ryota Tomioka, Taiji Suzuki, and Masashi Sugiyama, "Super-linear convergence of dual augmented lagrangian algorithm for sparsity regularized estimation," *Journal of Machine Learning Research*, vol. 12, no. May, pp. 1537–1586, 2011.