SPARSE ALGORITHMS AND BOUNDS FOR STATISTICALLY AND COMPUTATIONALLY EFFICIENT ROBUST ESTIMATION

S. Schuster

vatron gmbh
Linz, Stahlstr. 14, AT-4031, Linz, Austria
phone: +43(0)6648364397, email: stefan.schuster@vatron.com
web: www.vatron.com

ABSTRACT

Robust estimators that provide accurate parameter estimates even under the condition that classical assumptions like outlier-free additive Gaussian measurement noise do not hold exactly are of great practical importance in signal processing and measurement science in general. Lots of methods for deriving robust estimators exist. In this paper, we derive novel algorithms for robust estimation by modeling the outliers as a sparse additive vector of unknown deterministic or random parameters. By exploiting the separability of the estimation problem and applying recently developed sparse estimation techniques, algorithms that remove the effect of the outlying observations can be developed. Monte Carlo simulations show that the performance of the developed algorithms is practically equal to the best possible performance given by the Crámer-Rao lower bound (CRB) and the mean-squared error (MSE) of the oracle estimator [1], demonstrating the high accuracy. It is shown that the algorithms can be implemented in a computationally efficient manner. Furthermore, some interesting connections to the popular least absolute deviation (LAD) estimator are shown.

1. INTRODUCTION AND SIGNAL MODELS

The problem of estimating the vector of p unknown parameters $\theta = [\ \theta_1\ \ \theta_2\ \ \dots\ \ \theta_p\ ]^{\mathrm{T}}$ from the sampled data

$$x[n] = s(n,\theta) + w[n] \qquad n = 0,1,\dots,N-1 \tag{1}$$

has been extensively studied. N is the number of sampled data and $s(n,\theta)$ can be a linear or nonlinear signal model. Usually, the additive measurement noise $w[n]$ is assumed to be white Gaussian noise, or at least it is assumed that $w[n]$ is free of outliers, i.e., the observations are assumed to be free of gross errors[1]. Then, standard methods for deriving parameter estimators like the method of least squares (LS) can be applied. However, if the data contains outliers, the method of LS is known to fail completely, even if only a single outlier is present in the data [2, p. 11]. A bunch of robust estimators that suppress the disastrous effects of the outliers have been derived over the years, see, e.g., [2], [3]. Among them, popular robust estimators are the least median of squares (LMS) estimator and the least trimmed squares (LTS) estimator. They are especially designed to possess a high breakdown-point, i.e., they can handle are large number up to 50 percent of outliers, the best value possible. See [2] for a summary. Another well known method is the LAD estimator (also called least absolute value regression or L_1 regression estimator) [4]. The idea is to impose a Laplacian distribution for $w[n]$ instead of a Gaussian distribution. The Laplacian distribution is a heavy-tailed distribution and hence takes outlying observations into account. Effectively, the impact of outlying observations is reduced because their effect in the underlying L_1-norm cost function $J_{\mathrm{LAD}} = \|\mathbf{x} - \mathbf{H}\theta\|_1$ is de-emphasized compared to LS. Similarly, the seminal work of Huber [3] lead to a new class of estimators called M-estimators, which aim to minimize the impact of outliers by a modification of the cost function that down weights large residuals. Often, especially the LAD estimator is implemented by a variant of the popular iteratively reweighted LS (IRLS) algorithm, see, e.g., [4] and the recent paper [5]. It can be easily shown that the L_1-norm optimization problem $\min_{\theta} J_{\mathrm{LAD}}$ can be approximately solved by the iterative algorithm

$$\hat{\mathbf{r}}_{(k)} = \mathbf{x} - \mathbf{H}\hat{\theta}_{\mathrm{LAD},(k)}$$

$$\mathbf{W}_{\mathrm{LAD},(k)} = \mathrm{diag}\left\{ 1./\left|\hat{\mathbf{r}}_{(k)}\right| \right\} \tag{2}$$

$$\hat{\theta}_{\mathrm{LAD},(k+1)} = \left(\mathbf{H}^{\mathrm{T}}\mathbf{W}_{\mathrm{LAD},(k)}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{W}_{\mathrm{LAD},(k)}\mathbf{x}, \tag{3}$$

with an arbitrary initial estimate for $\hat{\theta}_{\mathrm{LAD},(1)}$. The subscript (k) denotes the kth iteration, $\mathrm{diag}\{\cdot\}$ denotes a diagonal matrix. The diagonal elements of $\mathbf{W}_{\mathrm{LAD},(k)}$ are calculated by the component wise reciprocal of the estimated residuals $\hat{\mathbf{r}}_{(k)}$, denoted by $1./\left|\hat{\mathbf{r}}_{(k)}\right|$. If, during the iteration, some of the residuals become zero or close to zero, they are temporarily replaced by some small but nonzero constant ε. The LAD estimator, while being very simple, has a relatively poor statistical performance [6, p. 48]. We will later discuss the reason therefor in detail. On the other hand, the LMS and LTS algorithm are computationally quite demanding and their statistical performance is also very poor. Usually, a second estimation step is employed to improve their accuracy, like it has been done in the MM estimation technique [6, p. 56]. In this work, we completely depart from above techniques and derive robust estimators by modifying the model (1). For simplicity, we limit ourselves to a linear signal model, i.e., $\mathbf{s}(\theta) = \mathbf{H}\theta$, with $\mathbf{H} \in \mathbb{R}^{\mathrm{Nxp}}$. However, the algorithms presented can be applied to nonlinear problems with straightforward modifications. Possible outliers are taken into account by adding an unknown sparse vector δ to (1), leading to the model (in matrix notation)

$$\mathbf{x} = \mathbf{H}\theta + \delta + \mathbf{w}. \tag{4}$$

Sparse means that the number of nonzero entries in $\delta \in \mathbb{R}^{\mathrm{Nx1}}$ is known to be (much) smaller than N. Furthermore, $\mathbf{w}$ is assumed to be uncorrelated zero mean Gaussian noise, i.e., $\mathbf{w} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$. To the best knowledge of the authors, the first who proposed such a sparse model for taking outliers into account were Mattingley and Boyd in the context of robust Kalman filtering [7]. For batch estimation, in a very recent paper Jin and Rao [8] proposed to use such a model. In the latter paper, Bayesian maximum a posteriori (MAP) and empirical Bayesian methods have been used to derive robust estimators. Here, we treat the case where the elements of δ are either purely deterministic or random variables. We will show that the different modeling assumptions lead to different algorithms, which have good statistical performance and are simple to be implemented. Based on some of our results, we outline interesting connections to the LAD estimator, explaining its relatively poor statistical performance. In fact, (4) seems to be a more realistic model

[1] In this work, we only treat outliers in $x[n]$. Outliers on the abscissa, so called leverage points, are not treated.

for outlier-contaminated data than, e.g., the modeling assumption that the noise is Laplacian distributed as it has been done for the LAD estimator or the "artificial" modifications of the cost function leading to the class of M-estimators. In practice, those data samples that are free of outliers can usually be well modeled by a deterministic variable plus measurement noise.

2. ROBUST ESTIMATORS AND PERFORMANCE BOUNDS

2.1 Deterministic Model

For the deterministic version of the model (4), a conceptually straightforward way to obtain estimates for θ and the nuisance parameters δ would be the application of the maximum likelihood (ML) technique, which leads to the optimization problem

$$\min \|\mathbf{x}-\mathbf{H}\theta-\delta\|_2^2 \text{ s.t. } \|\delta\|_0 \leq \alpha. \tag{5}$$

The Euclidean norm of a vector is denoted by $\|\cdot\|_2$, the semi-norm $\|\cdot\|_0$ denotes the number of nonzero entries of a vector, which is bounded by the constant α, and the abbreviation s.t. denotes subject to. However, above optimization problem is non-convex and NP-hard due to the used $\|\cdot\|_0$ semi-norm [1]. Hence, a computationally efficient algorithm for solving it is unlikely to exist. A key to derive simple estimators is to recognize the separability of the estimation problem into a sparse and a non-sparse part, enabling the application of a technique very similar to the principle of separable LS [9], [10]. Let $\mathbf{z}=\mathbf{x}-\delta$. Then, the probability density function (PDF) of $\mathbf{z}$, $p(\mathbf{z};\theta)$, is an outlier-free standard Gaussian distribution. On the other hand, for known θ and unknown δ, $\mathbf{x}-\mathbf{H}\theta=\delta+\mathbf{w}$ can be viewed as the problem of estimating a sparse deterministic vector, a problem which has received some attention in the recent literature [11]. We will next show different methods to exploit this.

2.1.1 Basis Pursuit Denoising

A possibility to derive a robust estimator based on the model (4) under the assumption that δ is a purely deterministic but sparse vector is the technique of basis pursuit denoising (BPDN) [12], [1]. Here, relaxing the $\|\cdot\|_0$ semi-norm by replacing it by the L_1-norm leads to the cost function

$$J_{\text{BP}}=\frac{1}{2}\|\mathbf{x}-\mathbf{H}\theta-\delta\|_2^2+\lambda\|\delta\|_1, \tag{6}$$

analogous to the MAP approach in the Bayesian case presented in [8]. The regularization parameter λ allows to adjust the expected sparsity. Equation (6) is a convex cost function and thus can be reliably numerically minimized by freely available solvers like `CVX`, a package for specifying and solving convex programs [13], [14]. However, instead of using a generic solver, we can gain further insight by deriving a very simple IRLS-like algorithm for solving (6). Let $\Phi(\delta)=\text{diag}\{|\delta|\}$ and attempt to approximate the minimization of (6) by the recursive minimization of

$$\begin{aligned} J_{\text{BP,IRLS}} \;=\; & \frac{1}{2}\left\|\underbrace{\mathbf{x}-\mathbf{H}\theta_{\text{BP},(k+1)}}_{\mathbf{y}_{(k+1)}}-\delta_{\text{BP},(k+1)}\right\|_2^2+ \\ & \lambda\delta_{\text{BP},(k+1)}^{\text{T}}\Phi_{(k)}^{-1}\delta_{\text{BP},(k+1)}. \end{aligned}$$

For notational convenience, the nonlinear dependence of $\Phi_{(k)}$ from $\delta_{\text{BP},(k)}$ has been suppressed. We now exploit the already mentioned separability of the problem and first minimize

$$\begin{aligned} J_{\text{BP,IRLS}} \;=\; & \frac{1}{2}\left(\mathbf{y}_{(k+1)}-\delta_{\text{BP},(k+1)}\right)^{\text{T}}\left(\mathbf{y}_{(k+1)}-\delta_{\text{BP},(k+1)}\right) \\ & +\lambda\delta_{\text{BP}(k+1)}^{\text{T}}\Phi_{(k)}^{-1}\delta_{\text{BP},(k+1)} \end{aligned} \tag{7}$$

with respect to $\delta_{\text{BP},(k+1)}$. This quadratic optimization problem has the closed-form solution

$$\hat{\delta}_{\text{BP},(k+1)}=\left(\mathbf{I}+2\lambda\Phi_{(k)}^{-1}\right)^{-1}\mathbf{y}_{(k+1)}, \tag{8}$$

where $\mathbf{I}$ denotes the identity matrix of appropriate dimensions. Using the abbreviation

$$\mathbf{A}_{(k)}=\left(\mathbf{I}+2\lambda\Phi_{(k)}^{-1}\right)^{-1}, \tag{9}$$

back substitution of (8) in (7) yields

$$\begin{aligned} J_{\text{BP,IRLS}}^{'} \;=\; & \frac{1}{2}\left(\mathbf{y}_{(k+1)}-\mathbf{A}_{(k)}\mathbf{y}_{(k+1)}\right)^{\text{T}}\left(\mathbf{y}_{(k+1)}-\mathbf{A}_{(k)}\mathbf{y}_{(k+1)}\right) \\ & +\;\lambda\mathbf{y}_{(k+1)}^{\text{T}}\mathbf{A}_{(k)}\Phi_{(k)}^{-1}\mathbf{A}_{(k)}\mathbf{y}_{(k+1)}. \end{aligned}$$

Expanding above equation, grouping common terms, and some further work results in

$$\begin{aligned} J_{\text{BP,IRLS}}^{'} \;=\; & \frac{1}{2}\mathbf{y}_{(k+1)}^{\text{T}}\mathbf{y}_{(k+1)}-\mathbf{y}_{(k+1)}^{\text{T}}\mathbf{A}_{(k)}\mathbf{y}_{(k+1)}+ \\ & \frac{1}{2}\mathbf{y}_{(k+1)}^{\text{T}}\mathbf{A}_{(k)}\mathbf{A}_{(k)}\mathbf{y}_{(k+1)}+\lambda\mathbf{y}_{(k+1)}^{\text{T}}\mathbf{A}_{(k)}\Phi_{(k)}^{-1}\mathbf{A}_{(k)}\mathbf{y}_{(k+1)} \\ \;=\; & \frac{1}{2}\mathbf{y}_{(k+1)}^{\text{T}}\mathbf{y}_{(k+1)}- \\ & \frac{1}{2}\mathbf{y}_{(k+1)}^{\text{T}}\left[2\mathbf{A}_{(k)}-\mathbf{A}_{(k)}\underbrace{\left(\mathbf{I}+2\lambda\Phi_{(k)}^{-1}\right)}_{\mathbf{A}_{(k)}^{-1}}\mathbf{A}_{(k)}\right]\mathbf{y}_{(k+1)} \\ \;=\; & \mathbf{y}_{(k+1)}^{\text{T}}\left(\mathbf{I}-\mathbf{A}_{(k)}\right)\mathbf{y}_{(k+1)} \end{aligned} \tag{10}$$

We now denote

$$\mathbf{W}_{\text{BP},(k)}=\mathbf{I}-\mathbf{A}_{(k)} \tag{11}$$

and use

$$\mathbf{y}_{(k+1)}=\mathbf{x}-\mathbf{H}\theta_{\text{BP},(k+1)} \tag{12}$$

in (10), obtaining

$$J_{\text{BP,IRLS}}^{''}=\left(\mathbf{x}-\mathbf{H}\theta_{\text{BP},(k+1)}\right)^{\text{T}}\mathbf{W}_{\text{BP},(k)}\left(\mathbf{x}-\mathbf{H}\theta_{\text{BP},(k+1)}\right) \tag{13}$$

with the solution

$$\hat{\theta}_{\text{BP},(k+1)}=\left(\mathbf{H}^{\text{T}}\mathbf{W}_{\text{BP},(k)}\mathbf{H}\right)^{-1}\mathbf{H}^{\text{T}}\mathbf{W}_{\text{BP},(k)}\mathbf{x}. \tag{14}$$

To summarize, the algorithm is initialized by a vector with nonzero but otherwise arbitrary elements $\delta_{(1)}$. Using (9), (11), and thereafter (14), $\mathbf{A}_{(k)}$, $\mathbf{W}_{\text{BP},(k)}$, and finally $\hat{\theta}_{\text{BP},(k+1)}$ can be calculated. Evaluating (12), we obtain $\mathbf{y}_{(k+1)}$, and a new iteration can be started using (8). Typically, the iteration is terminated if $\left\|\hat{\theta}_{(k+1)}-\hat{\theta}_{(k)}\right\|_2$ is smaller than a predefined threshold.

The similarity between the LAD iteration (3) and the BPDN iteration (14) is striking. The only difference is how the individual weighting matrices $\mathbf{W}_{\text{LAD},(k-1)}$ and $\mathbf{W}_{\text{BP},(k-1)}$ are calculated. In fact, the cost function (13) suggests the interpretation of a weighted LS criterion where the weighting coefficients are estimated iteratively from the data. Shrinkage helps to improve the estimates of these weighting coefficients, leading to improved statistical performance, as the simulation results in Section 3 show. For the choice of λ some rules of thumb are available in the literature [15, p. 92].

2.1.2 Hard Thresholding

By exploiting the separability of the estimation problem, it is also possible to derive a computationally feasible algorithm for approximately solving (5). Defining $\mathbf{y}_{(k)} = \mathbf{x} - \mathbf{H}\theta_{(k)}$, we first tackle the sparse estimation problem

$$\min \left\| \mathbf{y}_{(k)} - \delta_{\mathrm{HT},(k+1)} \right\|_2^2 \quad \text{s.t.} \quad \left\| \delta_{\mathrm{HT},(k+1)} \right\|_0 \leq \alpha$$

by hard-thresholding

$$\hat{\delta}_{\mathrm{HT},(k+1)}[n] = \begin{cases} y_{(k)}[n] & \text{if } y_{(k)}[n] \geq T(k) \\ 0 & \text{else} \end{cases} \tag{15}$$

see, e.g., [11], [15, chap. 5]. The parameter $T(k)$ is a threshold discussed shortly. In a next step, the solution to quadratic optimization problem $\min_{\theta_{\mathrm{HT},(k+1)}} \left\| \mathbf{x} - \hat{\delta}_{\mathrm{HT},(k+1)} - \mathbf{H}\theta_{\mathrm{HT},(k+1)} \right\|_2^2$ is given by

$$\hat{\theta}_{\mathrm{HT},(k+1)} = \left(\mathbf{H}^{\mathrm{T}}\mathbf{H} \right)^{-1} \mathbf{H}^{\mathrm{T}} \left(\mathbf{x} - \hat{\delta}_{\mathrm{HT},(k+1)} \right). \tag{16}$$

Thereafter, $\mathbf{y}_{(k+1)} = \mathbf{x} - \mathbf{H}\hat{\theta}_{\mathrm{HT},(k+1)}$ and a new iteration can begin. The iteration can be initialized with an arbitrary value for $\theta_{(1)}$. To turn the above procedure into a working algorithm, $T(k)$ must be adaptively adjusted. We found the adaptive threshold $T(k) = \hat{\sigma}_{(k)}\sqrt{2\log N}$ with

$$\hat{\sigma}_{(k)} = \mathrm{std}\left\{ \mathbf{y}_{(k)} - \hat{\delta}_{\mathrm{HT},(k)} \right\}$$

to work well, which is motivated by the commonly applied threshold $T = \sigma\sqrt{2\log N}$ for hard-thresholding based estimation of a sparse vector in Gaussian noise [11]. An even simpler alternative is to initialize the algorithm with the results of the LAD algorithm, somewhat similar to the idea how MM-estimators improve the performance of a prior estimation step with low efficiency. In this case, only one iteration (15) and (16) suffices. Simulation results in Section 3 show the good statistical properties of above described procedure.

2.2 Random Model

Instead of treating δ in (4) as deterministic, here we discuss the case that δ is a vector of zero-mean uncorrelated random variables with possibly different variances $\delta[n] \sim \mathcal{N}\left(0, \sigma_n^2\right)$. Furthermore, it is known that δ is sparse, i.e., σ_n^2 is zero for most n. However, we first intentionally neglect this assumed sparsity and attempt to find an estimator for the augmented vector of unknown parameters $\xi = \left[\begin{array}{cccc} \theta^{\mathrm{T}} & \sigma_0^2 & \sigma_1^2 & \ldots & \sigma_{N-1}^2 \end{array} \right]^{\mathrm{T}}$ of the signal model

$$\mathbf{x} = \mathbf{H}\theta + \delta + \mathbf{w}. \tag{17}$$

The PDF of $\delta + \mathbf{w}$ is Gaussian with a diagonal covariance matrix $\mathbf{C}$ depending on the unknown variances, $\delta + \mathbf{w} \sim \mathcal{N}\left(\mathbf{0}, \mathrm{diag}\left\{ \left[\begin{array}{cccc} \sigma^2 + \sigma_0^2, & \sigma^2 + \sigma_1^2, & \ldots, & \sigma^2 + \sigma_{N-1}^2 \end{array} \right]^{\mathrm{T}} \right\}\right)$. To derive an ML estimator, we have to solve [10, p. 185, p. 73]

$$\begin{aligned} \frac{\partial \log p(\mathbf{x};\xi)}{\xi_i} \;=\; & \frac{1}{2}\mathrm{tr}\left\{ \mathbf{C}^{-1}(\xi)\frac{\partial \mathbf{C}(\xi)}{\partial \xi_i} \right\} + \frac{\partial \mu(\xi)^{\mathrm{T}}}{\partial \xi_i}\mathbf{C}^{-1}(\xi)\left(\mathbf{x} - \mu(\xi)\right) \\ & + \frac{1}{2}\left(\mathbf{x} - \mu(\xi)\right)^{\mathrm{T}}\mathbf{C}^{-1}(\xi)\frac{\partial \mathbf{C}(\xi)}{\partial \xi_i}\left(\mathbf{x} - \mu(\xi)\right) = 0 \end{aligned}$$

for $i = 1, 2, \ldots, p+N$, $\mathrm{tr}\{\cdot\}$ denotes the sum of the diagonal elements of a matrix. Here, μ denotes the mean of the signal model. After some lengthy calculations which can not be presented due to the lack of space, it turns out that the following nonlinear system of equations

$$\mathbf{H}^{\mathrm{T}}\mathbf{W}_{\mathrm{RM}}\left(\hat{\theta}\right)\mathbf{x} = \mathbf{H}^{\mathrm{T}}\mathbf{W}_{\mathrm{RM}}\left(\hat{\theta}\right)\mathbf{H}\hat{\theta} \tag{18}$$

has to be solved to obtain an estimate of θ. The elements of the diagonal matrix $\mathbf{W}_{\mathrm{RM}}\left(\hat{\theta}\right)$ are then obtained as the component wise reciprocal of the estimated residuals $\hat{\mathbf{r}} = \mathbf{x} - \mathbf{H}\hat{\theta}$ squared:

$$[\mathbf{W}_{\mathrm{RM}}]_{n,n} = 1/\hat{r}[n]^2.$$

Note the similarity to the LAD method and to the method of BPDN presented earlier. An iterative solution of (18) is to calculate

$$\mathbf{W}_{\mathrm{RM},(k)} = \mathrm{diag}\left\{ \left[\begin{array}{cccc} 1/\hat{r}_{(k)}^2[0] & 1/\hat{r}_{(k)}^2[1] & \ldots & 1/\hat{r}_{(k)}^2[N-1] \end{array} \right] \right\} \tag{19}$$

$$\hat{\theta}_{\mathrm{RM},(k+1)} = \left(\mathbf{H}^{\mathrm{T}}\mathbf{W}_{\mathrm{RM},(k)}\mathbf{H} \right)^{-1} \mathbf{H}^{\mathrm{T}}\mathbf{W}_{\mathrm{RM},(k)}\mathbf{x}. \tag{20}$$

The only difference to the LAD method is that the residuals are squared, a direct result of the Gaussian model employed here. Above iteration suggests the interpretation of a weighted LS estimator where the weighting coefficients (i.e., the unknown variances σ_n^2) are estimated iteratively from the data. However, it is obvious that the method presented here can not be directly used. Although the derivation is based on the principle of ML, the desirable (asymptotic) properties of ML like unbiasedness and consistency are not guaranteed to hold by this estimator for the following reasons. Firstly, the number of unknown parameters is greater than the number of data samples N, and secondly, the number of unknowns grows with the number of samples. Both conditions are violations on the prerequisite assumptions that guarantee the asymptotically optimal properties of ML estimates [10]. To correct the situation, we have to include the additional information that δ is sparse. One possibility is to apply hard-thresholding to the residuals in every iteration, which would, however, necessitate to use an adaptive thresholding. A much easier method is to use the final residuals and estimates of a prior run of the LAD algorithm implemented through IRLS, and thereafter apply once hard-thresholding to the residuals with the threshold $T = \sigma\sqrt{2\log N}$. All values below T are set to σ. Thereafter, (19) and (20) are applied. If σ is unknown, it can be robustly estimated from the residuals, see, e.g., [2, pp. 202]. One application of (19) followed by (20) suffices to achieve good statistical performance, as simulation results in Section 3 show. This simple method produces only a minor increase of the computational load. The fact that simple thresholding of the residuals improves the estimator's performance once again highlights the fact that the relatively poor statistical performance of the LAD estimator is caused by the poor estimate of the elements of $\mathbf{W}_{\mathrm{LAD}}$ in the case of Gaussian measurement noise.

2.3 Performance Bounds

2.3.1 Deterministic Model

Usually, the statistical performance of robust estimators is measured in terms of their (asymptotic) efficiency [6, pp. 9]. Their performance is compared to the performance of the estimator achieving maximum efficiency under some assumptions, often the LS estimator in the outlier-free case. We depart slightly from this method here and adopt the notion of the oracle estimator [1]. In our deterministic setting (4), the oracle estimator is an estimator which "knows" the positions of the outliers, i.e., the support $\mathrm{supp}\{\delta\}$. Then, the oracle estimator is given by

$$\hat{\theta}_{\mathrm{o}} = \left(\mathbf{H}_{\mathrm{o}}^{\mathrm{T}}\mathbf{H}_{\mathrm{o}} \right)^{-1} \mathbf{H}_{\mathrm{o}}^{\mathrm{T}}\mathbf{x},$$

where $\mathbf{H}_{\mathrm{o}}$ is the submatrix constructed from the rows of $\mathbf{H}$ corresponding to the position of the samples of $\mathbf{x}$ that are outlier-free. The covariance matrix of the oracle estimator is given by

$$\mathbf{C}_{\mathrm{o}} = \sigma^2 \left(\mathbf{H}_{\mathrm{o}}^{\mathrm{T}}\mathbf{H}_{\mathrm{o}} \right)^{-1}.$$

Of course the oracle estimator can not be implemented in practice but we can compare the MSE of our estimators to this bound. Its interpretation/relation to a CRB for sparse estimation problems is further discussed in [1]. The advantage is that this bound describes the best possible performance achievable by any unbiased estimator for a given number of outliers.

2.3.2 *Random Model*

For the random model (17) we derive the corresponding CRB, again intentionally neglecting the sparsity of δ first. The elements of the Fisher information matrix in our case are given by [10, p. 47]

$$[\mathbf{I}(\theta)]_{ij} = \frac{\partial \mu(\xi)^{\mathrm{T}}}{\partial \xi_i} \mathbf{C}^{-1}(\xi) \frac{\partial \mu(\xi)}{\partial \xi_j} + \frac{1}{2} \mathrm{tr} \left\{ \mathbf{C}^{-1}(\xi) \frac{\partial \mu(\xi)}{\partial \xi_i} \mathbf{C}^{-1}(\xi) \frac{\partial \mu(\xi)}{\partial \xi_j} \right\},$$

$i, j = 1, 2, \ldots, p+N$. After some lengthy calculations, the Fisher information matrix can be shown to be

$$\mathbf{I}(\theta) = \begin{bmatrix} \left(\mathbf{H}^{\mathrm{T}} \mathbf{C}^{-1} \mathbf{H}\right) & \mathbf{0} \\ \mathbf{0} & \frac{1}{2} \mathbf{C}^2 \end{bmatrix}.$$

Note the block-diagonal structure of the matrix. By inverting the Fisher information matrix, the CRB can be obtained. The best possible covariance matrix $\mathbf{C}_{\hat{\theta}\hat{\theta}}$ for any unbiased estimator $\hat{\theta}$ in the sense of positive semidefiniteness of the matrix $\mathbf{C}_{\hat{\theta}\hat{\theta}} - \mathbf{I}^{-1}(\theta)$ is given by

$$\mathbf{C}_{\hat{\theta}\hat{\theta}} = \left(\mathbf{H}^{\mathrm{T}} \mathbf{C}^{-1} \mathbf{H}\right)^{-1}.$$

If outliers are present the corresponding entries in $\mathbf{C}^{-1}$ become very small compared to the entries corresponding to outlier-free samples. Effectively, samples which contain outliers are strongly down weighted and hence the bound for the oracle estimator derived in the prior subsection and above bound for the random model are quantitatively very similar. Without taking the sparsity of δ into account, the estimator derived in Subsection 2.2 would not achieve this bound for reasons that have already been discussed. However, it is interesting to see the similarity of the bounds of the oracle estimator and the bound for the random model, against which the performance of the estimators derived will be compared in the next section.

3. SIMULATION RESULTS

To validate the presented algorithms and formulas, we first performed Monte Carlo simulations with 10^3 trials for each signal-to-noise ratio (SNR) SNR $= 1/\sigma^2$ for a 2nd-order polynomial signal model $s(n,\theta) = \theta_1 + \theta_2 n + \theta_3 n^2 = -0.1 + 7n + 3n^2$ in the deterministic model (4) with $\theta = [\ \theta_1\ \ \theta_2\ \ \theta_3\ ]^{\mathrm{T}}$. In Fig. 1, the case of no outliers is shown. The results of all estimators except the LAD estimator practically achieve the CRB (in terms of root mean square error (RMSE)) or their performance is at least very near to it. Shown are, exemplarily, the results for the 3rd polynomial coefficient, however the results for the other parameters are comparable. The results of the LAD estimator and the performance prediction for the LAD estimator, see [16], are also in very good agreement. In Fig. 2 and Fig. 3, simulation results for the deterministic model with a different percentage of the number of outliers with respect to N are shown. For BPDN, $\lambda = 0.001$ has been chosen. This constant choice of λ gives good performance even over the wide range of SNRs simulated. For an even wider range of SNRs λ should be chosen dependent on σ^2 for optimum performance. Also the results of the HT algorithm, both the IRLS-initialized version and the version with the adaptively chosen threshold (not shown), are in very good agreement with the CRB. The performance loss of the LAD estimator can clearly be seen and is in good agreement with the prediction. Also, the CRB for the case of no outliers is shown as an indication of the increase of the CRB due to the outliers. Results for the random model are given in Fig. 4 for an outlier contamination of 20 percent. The outliers have been generated according to a Gaussian distribution with $\mathcal{N}\left(0, \sigma_n^2 = 10^{12}\right)$ and the positions of the outliers have been randomly varied across all samples. The CRB shown is the CRB averaged over the individual CRBs. It can be seen that the estimator derived is in very good agreement with the CRB, outperforming the LAD estimator. It should also be mentioned that the deterministic estimators perform very well in case of the random model and vice versa. Finally, in Fig. 5, a simulation for constant SNR of 10 and varying percentage of outlier contamination is shown, with 10^3 Monte Carlo trials for every step. The outliers have been generated according to a random model with $\mathcal{N}\left(-10^6, \sigma_n^2 = 10^{12}\right)$. The nonzero mean was intentionally chosen to test the robustness of the algorithms in such a situation. Both the HT estimator derived for the deterministic and the estimator derived for the random model have been applied and show nearly the same RMSE at low and moderate percentage of outlier contamination compared to the RMSE of the oracle estimator. The RMSE of the BDPN method is slightly higher, but however, better than the RMSE of the LAD method. Because the breakdown point of the HT algorithm with LAD initialization was higher than the breakdown point with adaptive thresholding in this simulation we concentrated on the former case. The performance loss of the LAD method can clearly be identified. We also performed simulations where the coefficients of the polynomial have been randomly varied and also varied the SNR. For all simulations, the new estimators performed well at low to moderate SNRs, being in good agreement with the corresponding CRBs.

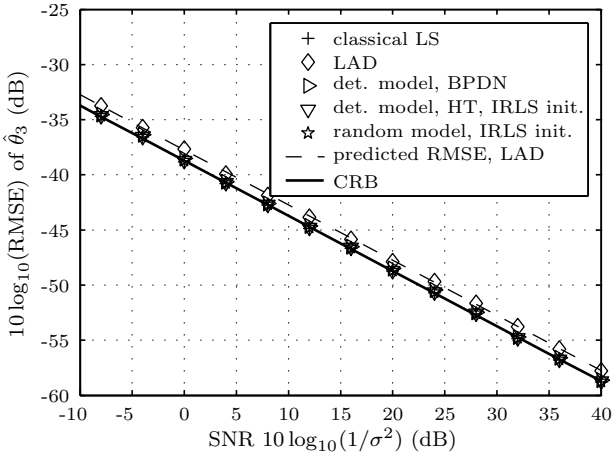

Figure 1: Simulation results for the case of no outlier. All estimators except the LAD are in very good agreement with the CRB.

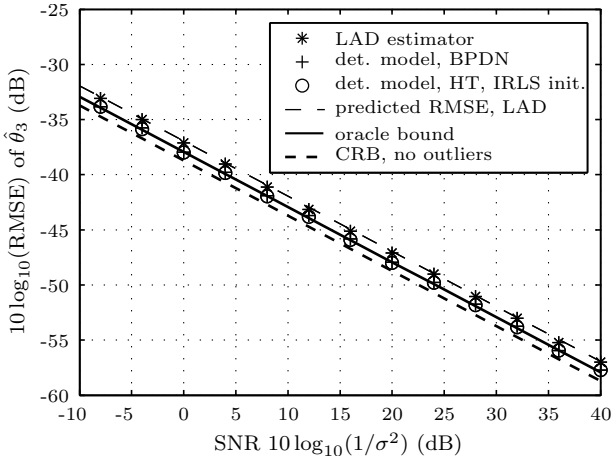

Figure 2: Simulation results for the deterministic model, 10 percent outlier contamination. The performance of all estimators except the LAD is in very good agreement with the oracle estimator bound.

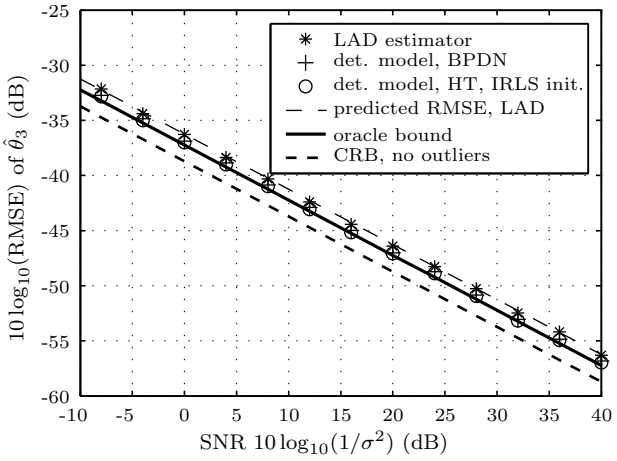

Figure 3: Simulation results for the deterministic model, 20 percent outlier contamination. The performance of all estimators except the LAD is in very good agreement with the oracle estimator bound.

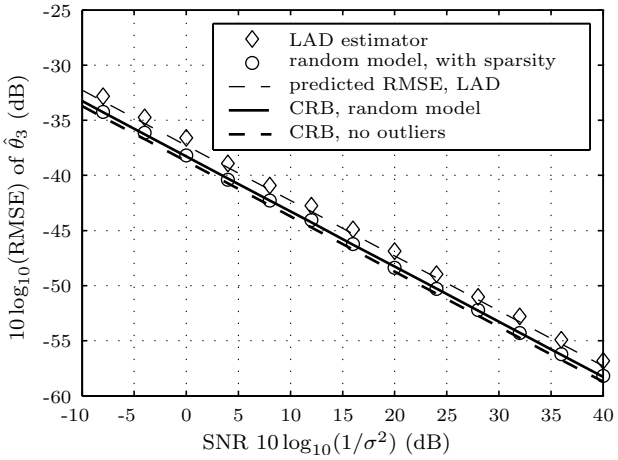

Figure 4: Simulation results for the random model, 20 percent outlier contamination. The performance of the estimator derived for the random model is in very good agreement with the CRB for the random model.

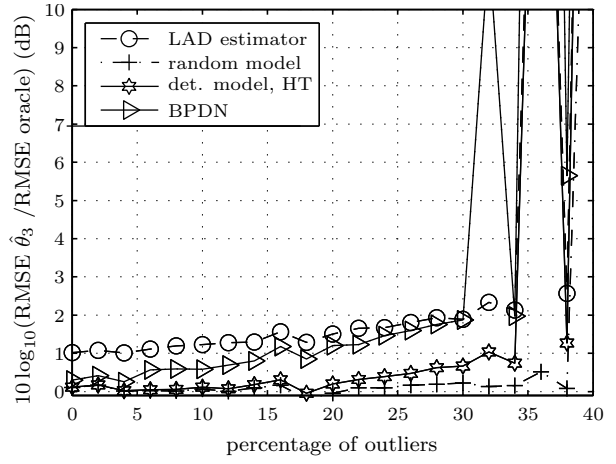

Figure 5: RMSE of different estimator compared to the RMSE of the oracle estimator in case of varying percentage of outlier contamination. Except the LAD estimator, at low to moderate contamination, the performance of the estimators is very close to the performance of the oracle estimator.

4. CONCLUSION

We presented new algorithms for robust estimation, which exploit a sparse model for outlying observations, together with the associated parameter estimation performance bounds. The results of Monte Carlo simulations indicate that the presented algorithms are in good agreement with the presented bounds, validating both the algorithms and the usefulness of the bounds. The presented algorithms can be easily implemented and are computationally quite efficient. Also, interesting differences/connections to LAD estimation are shown.

5. ACKNOWLEDGMENT

This work has been sponsored by the Austrian Center of Competence in Mechatronics (ACCM, www.accm.co.at) which is a COMET K2 center and is funded by the Austrian Federal Government, the Federal State Upper Austria, and its Scientific Partners.

REFERENCES

[1] Z. Ben-Haim and Y. C. Eldar, "The Cramér-Rao Bound for Estimating a Sparse Parameter Vector," *IEEE Trans. Signal Processing*, vol. 58, pp. 3384–3389, June 2010.

[2] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. Wiley-Interscience, 2003.

[3] P. J. Huber and E. M. Ronchetti, *Robust Statistics*. Wiley Interscience, 2009.

[4] G. A. F. Seber and C. J. Wild, *Nonlinear Regression*. New York, NY: Wiley Interscience, 2003.

[5] R. Chartrand and W. Yin, "Iteratively Reweighted Algorithms for Compressive Sensing," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2008, pp. 3869–3872.

[6] R. Andersen, *Modern Methods for Robust Regression*. SAGE Publications, 2008.

[7] J. Mattingley and S. Boyd, "Real-time convex optimization in signal processing," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 50–61, 2010.

[8] Y. Jin and B. D. Rao, "Algorithms for robust linear regression by exploiting the connection to sparse signal recovery," in *IEEE Acoustics, Speech, and Signal Processing Conf.*, 2010.

[9] G. H. Golub and V. Pereyra, "The Differentiation of Pseudoinverses and Nonlinear Least Squares Problems whose Variables Separate," *SIAM J. Numer. Anal.*, no. 10, pp. 413–432, 1973.

[10] S. M. Kay, *Fundamentals of Statistical Signal Processing-Estimation Theory*. Englewood Cliffs, Upper Saddle River, NJ: Prentice Hall, 1993.

[11] A. Jung, Z. Ben-Haim, F. Hlawatsch, and Y. C. Eldar, "On Unbiased Estimation Of Sparse Vectors Corrupted By Gaussian Noise," in *IEEE Acoustics, Speech, and Signal Processing Conf. (ICASSP 2010)*, 2010.

[12] J. A. Tropp, "Just relax: convex programming methods for identifying sparse signals in noise," vol. 52, no. 3, pp. 1030–1051, 2006.

[13] M. Grant and S. Boyd, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110, http://stanford.edu/ boyd/graph_dcp.html.

[14] ——, "{CVX}: Matlab software for disciplined convex programming, version 1.21," http://cvxr.com/cvx, Feb. 2011.

[15] M. Elad, *Sparse and Redundant Representations*. Springer-Verlag, 2010.

[16] T.-H. Li and K.-S. Song, "Estimation of the parameters of sinusoidal signals in non-gaussian noise," *IEEE Trans. Signal Processing*, vol. 57, pp. 62–72, 2009.