Robust Censoring Using Metropolis-Hastings Sampling

Georg Kail, *Member, IEEE*, Sundeep Prabhakar Chepuri, *Student Member, IEEE*, and Geert Leus, *Fellow, IEEE*

***Abstract*—The tasks of online data reduction and outlier rejection are both of high interest when large amounts of data are to be processed for inference. Rather than performing these tasks separately, we propose a joint approach, i.e., robust censoring. We formulate the problem as a non-convex optimization problem based on the data model for outlier-free data, without requiring prior model assumptions about the outlier perturbations. Moreover, our approach is general in that it is not restricted to any specific data model and does not rely on linearity, uncorrelated measurements, or additive Gaussian noise. For a given desired compression rate, the choice of the reduced dataset is optimal in the sense that it jointly maximizes the likelihood together with the inferred model parameters. An extension of the problem formulation allows for taking the average estimation performance into account in a hybrid optimality criterion. To solve the problem of robust censoring, we propose a Metropolis-Hastings sampler method that operates on small subsets of the data, thus limiting the computational complexity. As a practical example, the problem is specialized to the application of robust censoring for target localization. Simulation results confirm the superiority of the proposed method compared to other approaches.**

***Index Terms*—Big data, censoring, Markov chain Monte Carlo method, Metropolis-Hastings sampler, outlier rejection, robustness, sparse sensing.**

I. Introduction

In this era of data deluge, performing analytics on the massive volumes of data generated by ubiquitous sensors, internet, power and social networks, is increasingly challenging. Such prohibitively large volumes of datasets very often include entries from faulty systems, malicious agents, or entries that are irrelevant and redundant. The data generated from faulty systems, for example, may contain *outliers* that do not obey the postulated or learnt model [2]–[4]. These outlying entries significantly degrade the performance of the underlying inference tasks like prediction, estimation, tracking, clustering, and classification, to list a few. Therefore, the task of mining the most informative data samples and rejecting possible outliers is of paramount importance in data analytics.

Data analytics with large-scale data is infeasible without dimensionality reduction due to the limited computational capacity. Dimensionality reduction can be achieved by smartly designing efficient data gathering or sketching techniques keeping in mind the inference task to be performed, for example, through sensor selection or censoring. Sensor selection is an offline design approach [5]–[8], where the sensing operation is designed based only on the data model (even before gathering the data) such that a desired ensemble inference performance is achieved; hence, we refer to such methods as model-driven sensing schemes. On the other hand, in contrast to these offline design schemes, dimensionality reduction can be done on already acquired data by throwing away, i.e., censoring, less informative samples; we refer to such censoring schemes as data-driven sensing schemes. Censoring in its classical flavor is typically applied in a distributed setup, where the uninformative sensors do not transmit their observations to the fusion center [9], [10], thereby reducing the communications as well as the processing costs. Here, we do not assume a distributed setup and focus on the processing costs of subsequent inference tasks.

Evidently, the model-driven schemes that are agnostic to data are not robust to outliers. Even though the data-driven schemes like censoring use the data, existing state-of-the-art censoring schemes are not designed to be robust to outliers. On the other hand, existing robust estimators are not devised specifically for dimensionality reduction. Some well-known robust estimators are: (i) M-estimators [2], which are maximum likelihood-type estimators that replace the likelihood function with a smooth function introducing robustness, (ii) least-trimmed-squares (LTS) estimators [11], which remove outliers from the least-squares fit based on a predetermined breakdown point that determines the number of outliers, (iii) random sample consensus (RANSAC) [12], an iterative algorithm that classifies at each iteration a random subset of data as inliers or outliers, or (iv) sparsity-controlling outlier rejection (SCOR) [3], [13], [14], which models outliers as additive perturbations and estimates a sparse vector containing these perturbations. The above approaches reduce the dimensionality of the data only in so far as they discard outliers. They do not provide a meaningful tool for further dimensionality reduction. Moreover, they are mostly designed for linear Gaussian problems, without a straightforward generalization to more complicated non-linear or non-Gaussian estimation problems.

Manuscript received May 01, 2015; revised September 14, 2015; accepted November 20, 2015. Date of publication December 04, 2015; date of current version February 11, 2016. This work was supported in part by the Austrian Science Fund (FWF) under Grant J3495 and in part by NWO-STW under the VICI program (10382). A conference precursor of this manuscript appeared in the 16th IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC) July 2015 [1]. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Alfred Hero.

G. Kail is with the Institute of Telecommunications, Vienna University of Technology, 1040 Vienna, Austria (e-mail: georg.kail@tuwien.ac.at).

S. P. Chepuri and G. Leus are with the Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: s.p.chepuri@tudelft.nl; g.j.t.leus@tudelft.nl).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2015.2506142

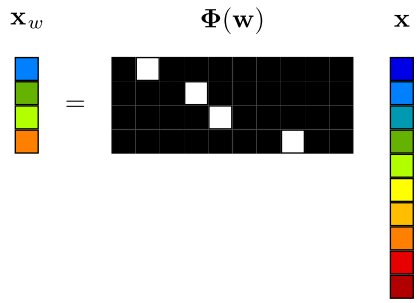

Fig. 1. Data censoring via sparse sensing, with $\mathbf{x}$, $\mathbf{\Phi}(\mathbf{w})$, and $\mathbf{x}_w$ denoting, respectively, the uncompressed data, the compression matrix, and the compressed data. A white, black, and colored square represent, respectively, a one, a zero, and an arbitrary value.

In this paper, we propose a joint approach for robust learning and data censoring, i.e., *robust censoring*. We focus on non-linear regression problems. The dimensionality reduction is achieved by linearly compressing the data through a sparse and deterministic compression matrix, that is, through *sparse sensing* [15] as shown in Fig. 1 (cf. [7], [16]).

We extend the existing literature on robust learning/estimation as well as censoring in various aspects:

- The proposed robust censoring theory is general, that is, it is not limited to an additive Gaussian noise model and can be applied to any data distribution. Furthermore, the data can be correlated.
- We do not assume a specific model for the outliers, as in many practical scenarios this model might not be known.
- In addition to a maximum likelihood robust censoring, we also propose a robust censoring scheme that includes the ensemble average estimation performance in its selection criterion.

We formulate the problem of robust censoring as a joint optimization problem that comprises compressive parameter estimation as well as the censoring mechanism. In contrast to compressive sensing [17], the parameter vector here need not be sparse. The proposed optimization amounts to selecting those measurements which fit the model of outlier-free data best. Since measurements containing outliers typically deviate from this model most, this approach ensures robustness with respect to outliers.

The resulting optimization problem is non-convex and non-linear. To solve it, we propose a method based on the powerful concept of Metropolis-Hastings (MH) sampling [18]–[22]. Like other Markov chain Monte Carlo (MCMC) methods [21], [22], MH sampling is a versatile iterative method capable of solving a wide range of problems that are too challenging for most classical estimation methods. It has been used, e.g., for classification [23], [24], for model learning [25], for uncertainty management [26], for random sampling from networks [27], [28], and as a complement for particle filters [29]–[31]. The flexibility of the MH concept requires that its implementation be carefully designed to ensure convergence within a moderate number of iterations. We present such a design for the robust censoring problem and subsequently further specialize it to provide a practical example.

Outline: This paper is organized as follows. Section II contains the problem formulation for robust censoring and relates it to existing schemes. In Section III, we present the proposed method for solving the problem using MH sampling. The general formulation of the method is subsequently specialized to the problem of *robust censoring for target localization* in Section IV, including numerical experiments to assess the performance of the proposed method. In Section V, we present a modification of the proposed robust censoring scheme that allows us to incorporate the average estimation performance into a hybrid optimality criterion. Our conclusions are summarized in Section VI.

Notation: The notation used in this paper can be described as follows. Upper (lower) bold face letters are used for matrices (column vectors). $(\cdot)^{\mathrm{T}}$ denotes transposition. $\mathrm{diag}(\cdot)$ refers to a diagonal matrix with its argument on the main diagonal. $\mathrm{diag}_{\mathrm{r}}(\cdot)$ represents a diagonal matrix with the argument on its diagonal but with the all-zero rows removed. $\mathbf{I}$ is an identity matrix. $\mathbb{E}\{\cdot\}$ denotes the expectation operation. The ℓ_0-(quasi) norm of a vector $\mathbf{w}$ refers to the number of nonzero elements in $\mathbf{w}$, i.e., $\|\mathbf{w}\|_0 := |\{m : \ w_m \neq 0\}|$. The ℓ_1-norm of an $N \times 1$ vector $\mathbf{w}$ is denoted by $\|\mathbf{w}\|_1 = \sum_{n=1}^{N} |w_n|$.

II. Problem Modeling

Consider a general non-linear regression problem, where an unknown vector $\boldsymbol{\theta} \in \mathbb{R}^N$ is to be estimated from the output data $\{x_m\}_{m=1}^{D}$. The output data are collected in the vector $\mathbf{x} = [x_1, x_2, \ldots, x_D]^{\mathrm{T}} \in \mathbb{R}^D$. We assume that the length-D data vector $\mathbf{x}$ is possibly contaminated with up to o outliers and/or it contains uninformative elements, where we interpret informative data as data that have a large likelihood.

Let the uncontaminated data vector, denoted by $\bar{\mathbf{x}} = [\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_D]^T \in \mathbb{R}^D$, be related to the unknown parameters $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_N]^{\mathrm{T}} \in \mathbb{R}^N$ through a non-linear model that is represented by the resulting probability distribution of $\bar{\mathbf{x}}$:

$$\bar{\mathbf{x}} \sim p(\bar{\mathbf{x}}; \boldsymbol{\theta}). \tag{1}$$

We assume that this distribution is known for any $\boldsymbol{\theta}$. However, due to the presence of outliers, the observed data $\mathbf{x}$ do not always obey the above model. For estimating the unknown parameter $\boldsymbol{\theta}$, we have access only to the (partly) contaminated data $\mathbf{x}$. The statistical dependence of $\mathbf{x}$ on $\bar{\mathbf{x}}$ or $\boldsymbol{\theta}$ is not assumed to be known. In other words, we do not assume a specific model for the outliers. We only use the implicit characteristic of outlier-contaminated measurements that they strongly deviate from the known model, i.e., they have a very small (or even zero) probability, given the true $\boldsymbol{\theta}$.

The dimensionality reduction of the observed data $\mathbf{x}$ is represented by the sparse Boolean vector $\mathbf{w} \in \{0,1\}^D$, where $w_m = 0$ indicates that x_m is considered outlying or is censored, and $\mathbf{w} = [w_1, w_2, \ldots, w_D]^{\mathrm{T}}$. Recall that D is also the length of $\mathbf{x}$. Out of the D elements of $\mathbf{w}$, only $d \ll D$ are nonzero. Using $\mathbf{w}$, we construct the censoring matrix $\mathbf{\Phi}(\mathbf{w}) \in \{0,1\}^{d \times D}$ as

$$\mathbf{\Phi}(\mathbf{w}) = \mathrm{diag}_{\mathrm{r}}(\mathbf{w}),$$

where $\mathrm{diag}_{\mathrm{r}}(\cdot)$ denotes a diagonal matrix with the argument on its diagonal, but with the all-zero rows removed. The resulting

matrix $\mathbf{\Phi}(\mathbf{w})$ is a fat sparse binary matrix with one nonzero element in each row and at most one nonzero element in each column. In Fig. 1, for example, $\mathbf{\Phi}(\mathbf{w})$ equals $\mathrm{diag}_{\mathrm{r}}(\mathbf{w})$ with $\mathbf{w} = [0,1,0,1,1,0,0,1,0,0]^{\mathrm{T}}$. By applying the linear compression operator $\mathbf{\Phi}(\mathbf{w})$ to the observed data vector $\mathbf{x}$, we obtain the compressed data vector

$$\mathbf{x}_w = \mathbf{\Phi}(\mathbf{w})\,\mathbf{x} = \mathrm{diag}_{\mathrm{r}}(\mathbf{w})\,\mathbf{x},$$

which is of length $d \ll D$. The reduced dimension data vector $\mathbf{x}_w$ is subsequently used to solve the inverse or learning problem. The corresponding unobserved uncontaminated $\bar{\mathbf{x}}_w = \mathbf{\Phi}(\mathbf{w})\,\bar{\mathbf{x}}$ follows a known pdf

$$\bar{\mathbf{x}}_w \sim p(\bar{\mathbf{x}}_w; \boldsymbol{\theta}, \mathbf{w}), \tag{2}$$

which we will use for robust estimation of $\boldsymbol{\theta}$ from $\mathbf{x}_w$.

In this paper, we pose the problem of designing a Boolean censoring vector $\mathbf{w}$ (and hence, a censoring matrix $\mathbf{\Phi}$) to jointly reject the outliers and compress the data in order to reduce the costs involved in solving the inverse problem. Formally, the robust censoring problem is stated as follows.

Problem (Robust Censoring): Given the data vector $\mathbf{x} \in \mathbb{R}^D$ which is related to the unknown $\boldsymbol{\theta} \in \mathbb{R}^N$ through a known non-linear data model but possibly contaminated with up to o outliers: (a) design the Boolean censoring vector $\mathbf{w} \in \{0,1\}^D$ that chooses $d \leq D-o$ data samples discarding possible outliers as well as censoring less informative samples (samples with smaller likelihood) and (b) use this reduced dimension data to compute an estimate of $\boldsymbol{\theta}$.

The difference of this formulation from classical censoring is, evidently, that the presence of outliers is explicitly accounted for. The difference from classical outlier rejection, on the other hand, is that d may be chosen much smaller than the number of apparently outlier-free data samples. Choosing a smaller d and working only with d-dimensional subvectors of $\mathbf{x}$ often leads to significant reductions of the computational cost. Moreover, since d in robust censoring is no longer determined by the number of outliers, the approach requires only very weak assumptions about the actual number of outliers. For large D and small d, the postulated $o \leq D - d$ indeed allows for a large range of o. No further assumptions about the outliers are made.

Mathematically, the robust censoring problem can be formulated as the following optimization problem

$$(\hat{\boldsymbol{\theta}}, \hat{\mathbf{w}}) = \underset{(\boldsymbol{\theta},\mathbf{w})\in\mathbb{R}^N\times\mathcal{W}}{\arg\max}\; p(\mathbf{x}_w; \boldsymbol{\theta}, \mathbf{w}), \tag{3}$$

where

$$\mathcal{W} = \left\{\mathbf{w} \in \{0,1\}^D \mid \|\mathbf{w}\|_0 = d\right\},$$

and the pdf $p(\mathbf{x}_w; \boldsymbol{\theta}, \mathbf{w})$ is in fact the likelihood function of $\bar{\mathbf{x}}_w$ as in (2), but with $\mathbf{x}_w$ inserted in place of the unobserved $\bar{\mathbf{x}}_w$. This means that for a fixed d, we are choosing $\boldsymbol{\theta}$ and $\mathbf{w}$ that fit the uncontaminated data model best in the maximum likelihood sense. In particular, fitting $\mathbf{w}$ optimally to the known data model of the *uncontaminated* data is an effective means to ensure robustness with respect to outliers, since measurements containing outliers typically deviate from this model the most. We underline that the formulation in (3) is general in the sense that it is not restricted to a specific data model and does not rely on assumptions such as linearity, uncorrelated noise, or additive Gaussian noise.

In many applications, the data vector $\mathbf{x}$ has some specific structure; it may, for example, consist of subvectors that correspond to measurements from different sensors. Depending on the particular setting, it may make sense to make a joint decision about selecting or rejecting an entire subvector, rather than processing each element individually. This can easily be achieved by adapting the definition of $\mathcal{W}$ such that it contains only vectors with a suitable structure. The corresponding modifications of robust censoring algorithms are straight-forward. For ease of exposition, the rest of this paper considers only the case where $\mathbf{x}$ is unstructured.

Relation of (3) to Sparsity-Controlling Outlier Rejection: Sparsity-controlling outlier rejection (SCOR) [3] is a state-of-the-art method for outlier rejection in linear Gaussian problems. In [13], it was shown to outperform random sample consensus (RANSAC) [12]. Here, we show that for additive linear Gaussian models our approach to robust censoring according to (3) in fact amounts to SCOR, up to some modifications, and to least-trimmed-squares (LTS) [11]. The proposed approach can thus be interpreted as a generalization of these state-of-the-art methods.

In the linear Gaussian data model, the relation between the outlier-free data vector $\bar{\mathbf{x}}$ and the unknown parameter vector $\boldsymbol{\theta}$ is

$$\bar{\mathbf{x}} = \mathbf{A}\boldsymbol{\theta} + \mathbf{n}, \tag{4}$$

with a known matrix $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_D]^{\mathrm{T}} \in \mathbb{R}^{D\times N}$ and a noise vector $\mathbf{n}$ from the distribution $\mathcal{N}(\mathbf{n}; \mathbf{0}, \sigma^2\mathbf{I})$, i.e., from a multivariate Gaussian distribution with mean $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$, where σ^2 is known. In this case, the likelihood function $p(\bar{\mathbf{x}}; \boldsymbol{\theta})$ is given as

$$p(\bar{\mathbf{x}}; \boldsymbol{\theta}) = \mathcal{N}(\bar{\mathbf{x}}; \mathbf{A}\boldsymbol{\theta}, \sigma^2\mathbf{I}),$$

and the likelihood function of the reduced dimension data $\bar{\mathbf{x}}_w$ is

$$p(\bar{\mathbf{x}}_w; \boldsymbol{\theta}, \mathbf{w}) = \mathcal{N}(\bar{\mathbf{x}}_w; \mathbf{A}_w\boldsymbol{\theta}, \sigma^2\mathbf{I}), \tag{5}$$

with $\mathbf{A}_w = \mathrm{diag}_{\mathrm{r}}(\mathbf{w})\mathbf{A}$. We assume that $d \geq N$ and $\mathbf{A}_w$ has full column rank for all $\mathbf{w} \in \mathcal{W}$. The proposed approach to robust censoring (cf. (3)) for this problem is

$$\begin{aligned}(\hat{\boldsymbol{\theta}}, \hat{\mathbf{w}}) &= \underset{(\boldsymbol{\theta},\mathbf{w})\in\mathbb{R}^N\times\mathcal{W}}{\arg\max}\; \mathcal{N}(\mathbf{x}_w; \mathbf{A}_w\boldsymbol{\theta}, \sigma^2\mathbf{I}) \\ &= \underset{(\boldsymbol{\theta},\mathbf{w})\in\mathbb{R}^N\times\mathcal{W}}{\arg\min}\; \|\mathbf{x}_w - \mathbf{A}_w\boldsymbol{\theta}\|^2 \\ &= \underset{(\boldsymbol{\theta},\mathbf{w})\in\mathbb{R}^N\times\mathcal{W}}{\arg\min}\; \sum_{m=1}^{D} w_m \left(x_m - \mathbf{a}_m^{\mathrm{T}}\boldsymbol{\theta}\right)^2,\end{aligned} \tag{6}$$

which for $d = D - o$ is equivalent to the LTS approach. More specifically, if we use the residuals $r_m(\boldsymbol{\theta}) = x_m - \mathbf{a}_m^{\mathrm{T}}\boldsymbol{\theta}$ and

let $r^2_{[m]}(\boldsymbol{\theta})$ denote the squared residuals in ascending order (for some $\boldsymbol{\theta}$), we can express $\hat{\boldsymbol{\theta}}$ from (6) as

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^N} \min_{\mathbf{w}\in\mathcal{W}} \sum_{m=1}^{D} w_m r_m^2(\boldsymbol{\theta}) \\ &= \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^N} \sum_{m=1}^{D-o} r^2_{[m]}(\boldsymbol{\theta}),\end{aligned}$$

which is the classical LTS formulation. The problem formulation for SCOR, on the other hand, is

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^N} \min_{\mathbf{o}\in\mathbb{R}^D} \|\mathbf{x} - \mathbf{A}\boldsymbol{\theta} - \mathbf{o}\|^2 + \lambda\|\mathbf{o}\|_1, \tag{7}$$

where $\mathbf{o} = [o_1, o_2, \ldots, o_D]^{\mathrm{T}}$ represents the additive outlier perturbations, and the regularization parameter λ controls the assumed number of outliers. In (7), the ℓ_1-norm is used to achieve sparsity of $\mathbf{o}$, i.e., as a more practical substitute for ℓ_0-regularization. If we replace the ℓ_1-norm by the ℓ_0-norm, there is some λ for which (7) can be written as

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^N} \min_{\mathbf{o}\in\mathbb{R}^D} \|\mathbf{x} - \mathbf{A}\boldsymbol{\theta} - \mathbf{o}\|^2 \\ &\quad \text{subject to } \|\mathbf{o}\|_0 = D - d \\ &= \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^N} \min_{\mathbf{o}\in\mathbb{R}^D} \sum_{m=1}^{D} \left(x_m - \mathbf{a}_m^{\mathrm{T}}\boldsymbol{\theta} - o_m\right)^2 \\ &\quad \text{subject to } \|\mathbf{o}\|_0 = D - d.\end{aligned} \tag{8}$$

Let us now define the binary vector $\mathbf{w} = [w_1, w_2, \ldots, w_D]^{\mathrm{T}}$ as

$$w_m = \begin{cases} 1 & \text{if } o_m = 0 \\ 0 & \text{else,} \end{cases}$$

which means that

$$o_m \in \mathcal{O}(w_m) = \begin{cases} \mathbb{R}\setminus\{0\} & \text{if } w_m = 0 \\ \{0\} & \text{else.} \end{cases} \tag{9}$$

Then, (8) can be written as

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^N} \min_{\mathbf{w}\in\mathcal{W}} \sum_{m=1}^{D} \min_{o_m\in\mathcal{O}(w_m)} \left(x_m - \mathbf{a}_m^{\mathrm{T}}\boldsymbol{\theta} - o_m\right)^2 \tag{10}$$

As is easily verified using (9), the summands in (10) simplify as follows:

$$\min_{o_m\in\mathcal{O}(w_m)} \left(x_m - \mathbf{a}_m^{\mathrm{T}}\boldsymbol{\theta} - o_m\right)^2 = w_m \left(x_m - \mathbf{a}_m^{\mathrm{T}}\boldsymbol{\theta}\right)^2,$$

and thus

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^N} \min_{\mathbf{w}\in\mathcal{W}} \sum_{m=1}^{D} w_m \left(x_m - \mathbf{a}_m^{\mathrm{T}}\boldsymbol{\theta}\right)^2,$$

which is identical to $\hat{\boldsymbol{\theta}}$ in (6). Thus, we have shown that if the ℓ_1-norm in SCOR is replaced by the ℓ_0-norm, there is some λ for which SCOR yields the same result as the proposed formulation of robust censoring.

III. Proposed Method: Metropolis-Hastings Sampling

Solving the problem of robust censoring as formulated in (3) is often too complex for direct calculation. In particular, it typically involves an exhaustive search over $\mathcal{W}$, which quickly becomes intractable for practical problem dimensions D and d. In this section, we present the proposed approach to solving (3) approximately. The proposed method relies on Bayesian sampling, more specifically Metropolis-Hastings (MH) sampling [18]–[22], which allows for a general formulation that requires only weak additional assumptions. To provide a practical example, the general formulation of the method in this section is subsequently specialized to the problem of robust censoring for target localization in Section IV and compared to other approaches through numerical experiments.

Straightforward Approach: Before we discuss the proposed method itself, we briefly sketch a straightforward deterministic approach to solving (3) approximately, namely by employing the alternating descent (AD) technique. This will later serve as a counter-example, motivating the use of the more complex but much more powerful MH methodology as proposed subsequently. The AD approach to solving (3) amounts to an iterative approximation by alternately fixing either $\boldsymbol{\theta}$ or $\mathbf{w}$ and maximizing $p(\mathbf{x}_w; \boldsymbol{\theta}, \mathbf{w})$ with regard to the respective other parameter. More specifically, in iteration i, we calculate

$$\boldsymbol{\theta}[i] = \arg\max_{\boldsymbol{\theta}\in\mathbb{R}^N} p\left(\mathbf{x}_w; \boldsymbol{\theta}, \mathbf{w}[i-1]\right) \tag{11}$$

$$\mathbf{w}[i] = \arg\max_{\mathbf{w}\in\mathcal{W}} p\left(\mathbf{x}_w; \boldsymbol{\theta}[i], \mathbf{w}\right). \tag{12}$$

A simple choice for the initialization is to use a $\mathbf{w}[0]$ that is randomly drawn from a uniform distribution over $\mathcal{W}$. As soon as maximization with respect to $\boldsymbol{\theta}$ or $\mathbf{w}$ does not change the respective parameter, a local maximum of $p(\mathbf{x}_w; \boldsymbol{\theta}, \mathbf{w})$ has been reached. The algorithm is thus terminated after iteration i if $\boldsymbol{\theta}[i] = \boldsymbol{\theta}[i-1]$ or $\mathbf{w}[i] = \mathbf{w}[i-1]$, returning $\hat{\boldsymbol{\theta}}_{\mathrm{AD}} = \boldsymbol{\theta}[i]$ and $\hat{\mathbf{w}}_{\mathrm{AD}} = \mathbf{w}[i]$.

Evidently, it depends on the shape of $p(\mathbf{x}_w; \boldsymbol{\theta}, \mathbf{w})$ whether the maxima in (11) and (12) can be calculated (possibly using some approximations). In general, (12) may again incur combinatorial complexity, no different from the original joint maximization in (3). However, in many problems fixing $\boldsymbol{\theta}$ or $\mathbf{w}$ simplifies the maximization of $p(\mathbf{x}_w; \boldsymbol{\theta}, \mathbf{w})$ with respect to the other parameter significantly. In particular, this is usually the case for the maximization with respect to $\mathbf{w}$ when the outlier-free observations $\bar{x}_m$ are statistically independent of each other (given $\boldsymbol{\theta}$), i.e., when $p(\bar{\mathbf{x}}; \boldsymbol{\theta})$ factorizes as $\prod_m p(\bar{x}_m; \boldsymbol{\theta})$. The main weakness of AD is that it converges only to a local maximum of $p(\mathbf{x}_w; \boldsymbol{\theta}, \mathbf{w})$ and is thus strongly influenced by the initialization. Simulations presented in Section IV-C confirm that AD is severely limited in the estimation performance it can achieve, compared to the proposed MH method, which will be presented next.

Target Distribution: Markov chain Monte Carlo (MCMC) methods [21], [22] such as MH sampling are iterative methods

that are often employed to calculate statistics of some probability distribution, the so-called *target distribution*[1], which may often be known only up to a normalization constant. Although the likelihood function $p(\mathbf{x}_w;\boldsymbol{\theta},\mathbf{w})$ in (3) is not a probability distribution of $\boldsymbol{\theta}$ and $\mathbf{w}$, it is non-negative and can thus be interpreted as a non-normalized probability distribution of $\boldsymbol{\theta}$ and $\mathbf{w}$. We could therefore directly use $p(\mathbf{x}_w;\boldsymbol{\theta},\mathbf{w})$ as the non-normalized target distribution and employ an MCMC method to maximize it with respect to $\boldsymbol{\theta}$ and $\mathbf{w}$. Here, however, we propose a slightly different approach. Using

$$\boldsymbol{\theta}_{\max}(\mathbf{w}) = \underset{\boldsymbol{\theta}\in\mathbb{R}^N}{\arg\max}\ p(\mathbf{x}_w;\boldsymbol{\theta},\mathbf{w}), \tag{13}$$

which is either calculated in closed form or obtained through some approximation, we can rewrite (3) as

$$\hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathcal{W}}{\arg\max}\ p\left(\mathbf{x}_w;\boldsymbol{\theta}_{\max}(\mathbf{w}),\mathbf{w}\right) \tag{14}$$

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_{\max}(\hat{\mathbf{w}}). \tag{15}$$

While $\hat{\mathbf{w}}$ and $\hat{\boldsymbol{\theta}}$ according to (14) and (15) are still the same as in (3), the advantage over (3) is that the formulation in (14), (15) allows us to use the target distribution

$$p_{\mathrm{t}}(\mathbf{w}) \propto p\left(\mathbf{x}_w;\boldsymbol{\theta}_{\max}(\mathbf{w}),\mathbf{w}\right), \tag{16}$$

which only needs to be maximized with respect to $\mathbf{w}$ rather than $(\boldsymbol{\theta},\mathbf{w})$. MCMC maximization over $\mathcal{W}$ is often simpler and faster than MCMC maximization over $\mathbb{R}^N \times \mathcal{W}$. After obtaining $\hat{\mathbf{w}}$ according to (14) by means of an MCMC method, we can calculate $\hat{\boldsymbol{\theta}}$ using (15).

Maximizing the target distribution $p_{\mathrm{t}}(\mathbf{w})$ according to the MCMC concept amounts to generating and processing a large population of realizations $\mathbf{w}^{(j)}$ from $p_{\mathrm{t}}(\mathbf{w})$. In the following, we will discuss how $\hat{\mathbf{w}}$ is estimated from the population and how the realizations are generated (since we cannot directly draw samples from $p_{\mathrm{t}}(\mathbf{w})$).

Sample-Based Estimation: Let J be the total number of realizations $\mathbf{w}^{(j)}$ used for maximizing $p_{\mathrm{t}}(\mathbf{w})$, and let $p_{\mathcal{S}}(\mathbf{w})$ denote the number of realizations $\mathbf{w}^{(j)}$ that are equal to the respective value of $\mathbf{w}$, normalized by J. Then, as J increases, $p_{\mathcal{S}}(\mathbf{w})$ tends to approximate $p_{\mathrm{t}}(\mathbf{w})$ more and more closely, as can easily be shown. The sample-based approximation of (14) is thus given by $\hat{\mathbf{w}}_{\mathcal{S}} = \arg\max_{\mathbf{w}\in\mathcal{W}}\ p_{\mathcal{S}}(\mathbf{w})$. However, as discussed in [32], [33] in more detail, for moderate sample sizes this approximation may be exceedingly coarse, which often makes the following widely-used alternative approach (see, e.g., [34]) preferable:

$$\hat{\mathbf{w}}_{\mathrm{eval}} = \mathbf{w}^{(j_{\max})} \quad \text{with}\ \ j_{\max} = \underset{j}{\arg\max}\ p_{\mathrm{t}}\left(\mathbf{w}^{(j)}\right). \tag{17}$$

[1]Contrary to the typical use of MCMC methods for Bayesian estimation, where the target distribution is a posterior probability distribution, we will not assign such a notion to the target distribution here. Nevertheless, our optimization problem according to (3) could in fact also be interpreted from a Bayesian perspective, namely as a maximum *a posteriori* (MAP) estimation with noninformative uniform priors on $\boldsymbol{\theta}$ and $\mathbf{w}$. The MH sampler that will be proposed in this section solves (3) and would thus perform MAP estimation, even though the target distribution will not be the corresponding posterior but merely have the same maximum as the posterior.

This approximation of $\hat{\mathbf{w}}$ is obtained by calculating $p_{\mathrm{t}}(\mathbf{w}^{(j)})$ for all j and picking the maximum. This approach is particularly well matched to the iterative nature of MCMC methods, where each iteration generates a new realization $\mathbf{w}^{(j)}$. By contrast to $\hat{\mathbf{w}}_{\mathcal{S}}$, finding $\hat{\mathbf{w}}_{\mathrm{eval}}$ does not require storing the entire population; instead, we can simply compare each new realization $\mathbf{w}^{(j)}$ to the realization that previously maximized the target distribution. Thus, throughout the entire algorithm only one realization needs to be stored. Moreover, in MH methods such as the one proposed here, $p_{\mathrm{t}}(\mathbf{w}^{(j)})$ is typically already calculated in the process of generating $\mathbf{w}^{(j)}$, which means that no further computations are needed for the comparison with the previous maximum.

In principle, (17) would not require that the realizations are generated from $p_{\mathrm{t}}(\mathbf{w})$. Any distribution could be used for generating the realizations, as long as its domain includes the maximizer of $p_{\mathrm{t}}(\mathbf{w})$ within $\mathcal{W}$ (which is $\hat{\mathbf{w}}$ according to (14)). Since we do not assume any prior information on which elements of $\mathcal{W}$ may be more likely or less likely to maximize $p_{\mathrm{t}}(\mathbf{w})$, another intuitive choice would be to generate the realizations simply from a uniform distribution over $\mathcal{W}$. However, generating the realizations from $p_{\mathrm{t}}(\mathbf{w})$ ensures that a realization $\mathbf{w}^{(j)}$ is more likely to be equal to $\hat{\mathbf{w}}$ than to any other $\mathbf{w} \in \mathcal{W}$. Assuming the ideal case that the realizations are independent from each other, it can easily be shown that this increases the probability that even a moderate-sized set of realizations contains $\hat{\mathbf{w}}$, compared to the choice of a uniform distribution.

MH Sampling: As mentioned above, each MCMC iteration generates—if we ignore the transient influence of the initialization—one new random realization $\mathbf{w}^{(j)}$ from the target distribution $p_{\mathrm{t}}(\mathbf{w})$. This randomness is a fundamental difference between MCMC methods and other widely used estimation methods such as the expectation-maximization [35] or belief propagation [36] algorithms. The MCMC method that we propose to use here is MH sampling, which amounts to the following procedure. At iteration j, we first generate a proposal $\widetilde{\mathbf{w}}$ from some proposal distribution $q(\widetilde{\mathbf{w}}|\mathbf{w}^{(j-1)})$, whose shape depends on the realization from the previous iteration, i.e., $\mathbf{w}^{(j-1)}$. Then, the new realization $\mathbf{w}^{(j)}$ is obtained as

$$\mathbf{w}^{(j)} = \begin{cases} \widetilde{\mathbf{w}} & \text{with probability } \alpha_j \\ \mathbf{w}^{(j-1)} & \text{with probability } 1-\alpha_j, \end{cases} \tag{18}$$

where

$$\alpha_j = \min\left\{\frac{p_{\mathrm{t}}(\widetilde{\mathbf{w}})\,q\left(\mathbf{w}^{(j-1)}\,|\widetilde{\mathbf{w}}\right)}{p_{\mathrm{t}}\left(\mathbf{w}^{(j-1)}\right)\,q\left(\widetilde{\mathbf{w}}\,\middle|\mathbf{w}^{(j-1)}\right)}, 1\right\}. \tag{19}$$

Note that iterations where $\mathbf{w}^{(j)} = \mathbf{w}^{(j-1)}$ do not influence the estimate $\hat{\mathbf{w}}_{\mathrm{eval}}$ and thus constitute a futile computational overhead. It is therefore advantageous if $q(\cdot|\cdot)$ can be designed such that α_j is typically large, thus reducing the number of such futile iterations.

Since there is no simple relation between the number of iterations J and the estimation performance, we propose to predetermine the number of iterations J based on training. Other approaches for choosing J include, e.g., assessing when the distribution of the realizations has converged to a stationary distribution [37]. Evidently, a smaller J is sufficient if we are mostly interested in outlier rejection and less in optimality, whereas a

larger J is needed to find the exact maximum likelihood solution of (3). With a smaller J, the algorithm is less likely to use all observations; instead, it typically uses only a smaller part of them. This may be a desired effect of complexity reduction when optimality is less important.

Compared to the widely used method of Gibbs sampling [21], [38], [39], MH sampling is more general, since the proposal distribution $q(\cdot|\cdot)$ is not specified by the MH concept but can be chosen freely, under some mild conditions. In Gibbs sampling (and its variations such as the recently proposed partially collapsed Gibbs sampling [40]–[42]), the proposal vector $\widetilde{\mathbf{w}}$ at some iteration j is determined to be equal to $\mathbf{w}^{(j-1)}$ in some of its elements while other elements are randomly chosen using particular distributions. These distributions are derived from $p_{\mathrm{t}}(\mathbf{w})$ and entail that α_j is always 1, which is potentially advantageous, as explained above. The main reason why we choose MH sampling rather than Gibbs sampling here is the following. According to the Gibbs sampling concept, the set of elements of $\widetilde{\mathbf{w}}$ that are randomly sampled—which is typically only one element—is chosen based only on j (in a periodic manner) but not on $\mathbf{w}^{(j-1)}$. In MH sampling, we can choose the set smartly using $\mathbf{w}^{(j-1)}$, which may reduce the number of iterations needed by the algorithm, especially when D is large. Furthermore, in the most straight-forward Gibbs sampler implementation, the distribution for sampling some elements of $\widetilde{\mathbf{w}}$ is defined on $\{0,1\}^K$, where K is the number of those elements. This may very often lead to a new realization that is not an element of $\mathcal{W}$ and is thus not considered in (17). More sophisticated Gibbs sampler designs can avoid this effect, but they are typically computationally complex. In MH sampling, we can easily design $q(\cdot|\cdot)$ such that $\widetilde{\mathbf{w}}$ is always within $\mathcal{W}$.

As shown in the previous paragraph, the design of $q(\cdot|\cdot)$ influences the number of iterations needed by the algorithm. In fact, $q(\cdot|\cdot)$ critically determines the rate at which the estimators improve with increasing numbers of iterations. If $q(\widetilde{\mathbf{w}}|\mathbf{w}^{(j-1)})$ is concentrated around $\mathbf{w}^{(j-1)}$ too strongly, for example, the estimators may improve steadily but very slowly, and the algorithm may spend many iterations around local maxima of the target distribution. On the other hand, if $q(\widetilde{\mathbf{w}}|\mathbf{w}^{(j-1)})$ is completely independent of $\mathbf{w}^{(j-1)}$, it typically produces many proposals that correspond to small values of α_j and thus fail to improve the estimators because of (18).

Proposed Implementation: In view of the trade-off discussed in the previous paragraph, we use two different types of proposals, which we call "small-step" proposals and "large-step" proposals (cf., e.g., [23]). In each sampler iteration, we decide to make either, with some fixed small probability β, a "large-step" proposal or, with probability $1-\beta$, a "small-step" proposal. We can thus express $q(\widetilde{\mathbf{w}}|\mathbf{w}^{(j-1)})$ as

$$q\left(\widetilde{\mathbf{w}}\,\middle|\,\mathbf{w}^{(j-1)}\right) = \beta\, q_{\mathrm{large}}\left(\widetilde{\mathbf{w}}\,\middle|\,\mathbf{w}^{(j-1)}\right) + (1-\beta)\, q_{\mathrm{small}}\left(\widetilde{\mathbf{w}}\,\middle|\,\mathbf{w}^{(j-1)}\right). \quad (20)$$

"Small-step" proposals introduce small variations based on the respective previous realization $\mathbf{w}^{(j-1)}$. Since the probability distribution of $\mathbf{w}^{(j-1)}$ is $p_{\mathrm{t}}(\mathbf{w})$, the proposal $\widetilde{\mathbf{w}}$ is also generally likely to be in a region where $p_{\mathrm{t}}(\mathbf{w})$ is large. Furthermore, we may be able to exploit the local shape of $p_{\mathrm{t}}(\mathbf{w})$ around $\mathbf{w}^{(j-1)}$ to make proposals that are likely to improve $p_{\mathrm{t}}(\mathbf{w})$. "Large-step" proposals, on the other hand, are intended to reduce the risk of finding only a local maximum of the target distribution. To this end, they are chosen independently of $\mathbf{w}^{(j-1)}$. Beyond these considerations, our particular choices for the "small-step" and "large-step" proposal distributions are rather arbitrary.

A "small-step" proposal $\widetilde{\mathbf{w}}$ is obtained from $\mathbf{w}^{(j-1)}$ by changing the $m^{(\mathrm{add})}$-th element from 0 to 1 and the $m^{(\mathrm{rem})}$-th element from 1 to 0, where $m^{(\mathrm{add})}$ and $m^{(\mathrm{rem})}$ are randomly picked. We can thus write

$$q_{\mathrm{small}}\left(\widetilde{\mathbf{w}}|\mathbf{w}^{(j-1)}\right) = p_{\mathrm{add}}\left(m^{(\mathrm{add})}\,\middle|\,\mathbf{w}^{(j-1)}\right) \times p_{\mathrm{rem}}\left(m^{(\mathrm{rem})}\,\middle|\,\mathbf{w}^{(j-1)}\right), \quad (21)$$

for all $\widetilde{\mathbf{w}} \in \mathcal{W}$ that differ from $\mathbf{w}^{(j-1)}$ in two elements, and 0 for all other $\widetilde{\mathbf{w}}$. As a straight-forward choice, we propose to draw $m^{(\mathrm{add})}$ from a uniform distribution over all the zero elements of $\mathbf{w}^{(j-1)}$:

$$p_{\mathrm{add}}\left(m\,\middle|\,\mathbf{w}^{(j-1)}\right) = \frac{1}{D-d}, \quad (22)$$

for any of the $D-d$ indices $m \in \{1,2,\ldots,D\}$ such that $w_m^{(j-1)} = 0$. The analogous distribution for $m^{(\mathrm{rem})}$ would be

$$p_{\mathrm{rem}}\left(m\,\middle|\,\mathbf{w}^{(j-1)}\right) = \frac{1}{d},$$

for any of the d indices $m \in \{1,2,\ldots,D\}$ such that $w_m^{(j-1)} = 1$. However, we can reduce the number of iterations needed by the algorithm if we instead design $p_{\mathrm{rem}}(m|\mathbf{w}^{(j-1)})$ such that the probability of obtaining a $\widetilde{\mathbf{w}}$ with higher $p_{\mathrm{t}}(\widetilde{\mathbf{w}})$ is increased. The design of such a more advantageous distribution $p_{\mathrm{rem}}(m|\mathbf{w}^{(j-1)})$ depends on the specific shape of $p_{\mathrm{t}}(\mathbf{w})$ in a given problem. If, for example, some simplifying approximations allow us to factorize

$$p_{\mathrm{t}}(\mathbf{w}) \approx \prod_{m=1}^{D} p_m(w_m)$$

such that each factor $p_m(w_m)$ depends only on one element w_m, then we can simply choose

$$p_{\mathrm{rem}}\left(m\,\middle|\,\mathbf{w}^{(j-1)}\right) = g\left(\frac{p_m(1)}{p_m(0)}\right), \quad (23)$$

for any of the d indices $m \in \{1,2,\ldots,D\}$ such that $w_m^{(j-1)} = 1$. Here, $g(\cdot)$ is some normalized decreasing function. Indeed, it is easily verified that choosing $m^{(\mathrm{rem})}$ among the indices with smaller $p_m(1)/p_m(0)$ leads to a larger (approximate) $p_{\mathrm{t}}(\mathbf{w})$. Two examples for this approach will be given in Sections IV-C and V-B. Note that, following the same rationale, we can also choose $p_{\mathrm{add}}(m|\mathbf{w}^{(j-1)})$ as an increasing function of $p_m(1)/p_m(0)$. However, in view of the much larger domain of $p_{\mathrm{add}}(m|\mathbf{w}^{(j-1)})$ (comprising $D-d$ elements rather than d) we prefer the simpler uniform distribution (22) here.

"Large-step" proposals $\widetilde{\mathbf{w}}$ are chosen from a uniform distribution over $\mathcal{W}$, independently of $\mathbf{w}^{(j-1)}$ or $\mathbf{x}$:

$$q_{\mathrm{large}}\left(\widetilde{\mathbf{w}}|\mathbf{w}^{(j-1)}\right) = q_{\mathrm{large}}(\widetilde{\mathbf{w}}) = \frac{1}{\binom{D}{d}}. \quad (24)$$

Inserting (20)–(22) and (24) into (19) and using

$$c = \frac{\beta(D-d)}{(1-\beta)\binom{D}{d}},$$

we obtain

$$\alpha_j = \min\left\{\frac{p_{\mathrm{t}}(\widetilde{\mathbf{w}})\left(c + p_{\mathrm{rem}}\left(m^{(\mathrm{add})} \mid \widetilde{\mathbf{w}}\right)\right)}{p_{\mathrm{t}}\left(\mathbf{w}^{(j-1)}\right)\left(c + p_{\mathrm{rem}}\left(m^{(\mathrm{rem})} \mid \mathbf{w}^{(j-1)}\right)\right)}, 1\right\}, \tag{25}$$

if $\widetilde{\mathbf{w}} \in \mathcal{W}$ differs from $\mathbf{w}^{(j-1)}$ in two elements—where we denote the index of the new nonzero element by $m^{(\mathrm{add})}$ and the index of the new zero element by $m^{(\mathrm{rem})}$—and

$$\alpha_j = \min\left\{\frac{p_{\mathrm{t}}(\widetilde{\mathbf{w}})}{p_{\mathrm{t}}\left(\mathbf{w}^{(j-1)}\right)}, 1\right\}, \tag{26}$$

for all other $\widetilde{\mathbf{w}} \in \mathcal{W}$.

The MH sampler for robust censoring is summarized as Algorithm 1.

Algorithm 1 MH sampler for robust censoring

Initialize with any $\mathbf{w}^{(0)}$ from $\mathcal{W}$, $\hat{\mathbf{w}}_{\mathrm{eval}} \leftarrow \mathbf{w}^{(0)}$
Iterate for $j = 1, 2, \ldots, J$:
  With probability β (e.g., $\beta = 0.005$):
    Generate $\widetilde{\mathbf{w}}$ from (24) ("large-step")
  In the converse case:
    Generate $m^{(\mathrm{add})}$ from (22) and $m^{(\mathrm{rem})}$ from (23) and calculate $\widetilde{\mathbf{w}}$ ("small-step")
  Calculate α_j according to (25), (26)
  With probability α_j : $\mathbf{w}^{(j)} \leftarrow \widetilde{\mathbf{w}}$
  In the converse case: $\mathbf{w}^{(j)} \leftarrow \mathbf{w}^{(j-1)}$
  If $p(\mathbf{w}^{(j)}) > p(\hat{\mathbf{w}}_{\mathrm{eval}})$
    $\hat{\mathbf{w}}_{\mathrm{eval}} \leftarrow \mathbf{w}^{(j)}$

IV. Example: Robust Censoring for Target Localization

In this section, we apply the proposed method to the problem of target localization as a practical example for non-linear inverse problems where robust censoring can be useful. Sections IV-A and IV-B contain the signal model and the formulation of the proposed MH method for this problem, respectively. Section IV-C presents numerical results that assess the performance of the method.

A. Signal Model

We assume a setting where a large number of sensors at known positions attempt to localize several targets using noisy distance measurements [43]–[45]. The distance measurements are potentially contaminated with outliers. In the absence of perturbations such as noise and outliers, a very small arbitrary subset of the distance measurements would be enough for locating a target. Outlier contamination but also model-consistent noise typically lead to a large variation in the quality of the different distance measurements. It therefore makes sense to search for a relatively small subset of measurements that potentially lead to optimal localization. Furthermore, besides leading to a reliable estimate, robust censoring reduces the signal processing cost associated with the number of measurements that are used and/or stored.

We assume that localization is performed independently for each target. For the sake of simplicity, we thus consider only one target in the remainder of this section while the total number of available sensors D is still assumed to be large. In the present problem, $\boldsymbol{\theta}$ is the unknown position of the target in N dimensions, i.e., typically, $N = 2$ or $N = 3$. We denote the known positions of the D sensors as $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_D$. The noisy (outlier-free) distance measurement from the m-th sensor is

$$\bar{x}_m = \|\boldsymbol{\theta} - \mathbf{z}_m\| + n_m, \tag{27}$$

where the additive noise n_m is zero-mean Gaussian with variance σ_m^2. Since the level of uncertainty of distance measurements is often higher for larger distances, we assume that σ_m^2 increases with $\|\boldsymbol{\theta} - \mathbf{z}_m\|$ as

$$\sigma_m^2 = \sigma_0^2\left(\|\boldsymbol{\theta} - \mathbf{z}_m\| + \zeta\right)^{\gamma}, \tag{28}$$

with some known non-negative constants σ_0^2, ζ, and γ that depend on how the distance measurements x_m were obtained. Moreover, we assume that the noise at different sensors may in general be correlated (e.g., due to some sources of interference that affect several sensors):

$$\mathrm{E}\{n_k n_l\} = \rho_{kl}\sigma_k\sigma_l, \tag{29}$$

where the coefficients ρ_{kl} are known and $\rho_{kk} = 1$ for $k = 1, 2, \ldots, D$. Thus, the noise covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathrm{E}\{\mathbf{n}\mathbf{n}^{\mathrm{T}}\}$ with $\mathbf{n} = [n_1, n_2, \ldots, n_D]^{\mathrm{T}}$ depends on $\boldsymbol{\theta}$ through (28) and (29). Using $h_m(\boldsymbol{\theta}) = \|\boldsymbol{\theta} - \mathbf{z}_m\|$ and $\mathbf{h}(\boldsymbol{\theta}) = [h_1(\boldsymbol{\theta}), h_2(\boldsymbol{\theta}), \ldots, h_D(\boldsymbol{\theta})]^{\mathrm{T}}$, we can write the likelihood function of the problem as

$$p(\bar{\mathbf{x}}; \boldsymbol{\theta}) = \mathcal{N}\left(\bar{\mathbf{x}}; \mathbf{h}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta})\right).$$

The likelihood function of the reduced dimension data $\bar{\mathbf{x}}_w$ is thus

$$p(\bar{\mathbf{x}}_w; \boldsymbol{\theta}, \mathbf{w}) = \mathcal{N}\left(\bar{\mathbf{x}}_w; \mathbf{h}_w(\boldsymbol{\theta}, \mathbf{w}), \boldsymbol{\Sigma}_w(\boldsymbol{\theta}, \mathbf{w})\right), \tag{30}$$

with

$$\mathbf{h}_w(\boldsymbol{\theta}, \mathbf{w}) = \mathrm{diag}_{\mathrm{r}}(\mathbf{w})\mathbf{h}(\boldsymbol{\theta})$$
$$\boldsymbol{\Sigma}_w(\boldsymbol{\theta}, \mathbf{w}) = \mathrm{diag}_{\mathrm{r}}(\mathbf{w})\boldsymbol{\Sigma}(\boldsymbol{\theta})\left(\mathrm{diag}_{\mathrm{r}}(\mathbf{w})\right)^{\mathrm{T}}.$$

Consequently, the proposed approach to robust censoring (cf. (3)) for this problem is

$$\begin{aligned}(\hat{\boldsymbol{\theta}}, \hat{\mathbf{w}}) &= \underset{(\boldsymbol{\theta}, \mathbf{w}) \in \mathbb{R}^N \times \mathcal{W}}{\arg\max}\ \mathcal{N}\left(\mathbf{x}_w; \mathbf{h}_w(\boldsymbol{\theta}, \mathbf{w}), \boldsymbol{\Sigma}_w(\boldsymbol{\theta}, \mathbf{w})\right) \\ &= \underset{(\boldsymbol{\theta}, \mathbf{w}) \in \mathbb{R}^N \times \mathcal{W}}{\arg\min} \left\|\left(\boldsymbol{\Sigma}_w(\boldsymbol{\theta}, \mathbf{w})\right)^{-1/2}\left(\mathbf{x}_w - \mathbf{h}_w(\boldsymbol{\theta}, \mathbf{w})\right)\right\|^2 \\ &\qquad + \log\left|\boldsymbol{\Sigma}_w(\boldsymbol{\theta}, \mathbf{w})\right|, \end{aligned} \tag{31}$$

where $|\mathbf{C}|$ denotes the determinant of the matrix $\mathbf{C}$ and $\mathbf{C}^{-1/2}$ can be obtained from $\mathbf{C}^{-1}$ by Cholesky factorization. Analogously, inserting (30) into (13) yields

$$\begin{aligned}\boldsymbol{\theta}_{\max}(\mathbf{w}) = \underset{\boldsymbol{\theta} \in \mathbb{R}^N}{\arg\min} &\left\|\left(\boldsymbol{\Sigma}_w(\boldsymbol{\theta}, \mathbf{w})\right)^{-1/2}\left(\mathbf{x}_w - \mathbf{h}_w(\boldsymbol{\theta}, \mathbf{w})\right)\right\|^2 \\ &+ \log\left|\boldsymbol{\Sigma}_w(\boldsymbol{\theta}, \mathbf{w})\right|. \end{aligned} \tag{32}$$

Since there is no closed-form solution for (32), we resort to the following iterative approximation. Starting from some random initial $\boldsymbol{\theta}[0]$, we calculate updates $\boldsymbol{\theta}[i]$ for $i = 1, 2, \ldots, I$ by evaluating the right hand side of (32) with $\boldsymbol{\Sigma}_w(\boldsymbol{\theta}, \mathbf{w})$ consistently replaced by $\boldsymbol{\Sigma}_w(\boldsymbol{\theta}[i-1], \mathbf{w})$. Each update amounts to solving a non-linear least-squares problem, which we do approximately using the Gauss-Newton algorithm [46]. In our experiments, $I = 20$ updates were enough to make the influence of the initialization negligible. The resulting $\boldsymbol{\theta}[I]$ serves as an approximation to (32) and will be denoted as $\tilde{\boldsymbol{\theta}}_{\max}(\mathbf{w})$ in the rest of this section.

B. Proposed Method: MH Sampling

For the proposed MH sampler, we obtain the target distribution (cf. (16), (30))

$$p_{\mathrm{t}}(\mathbf{w}) \propto \frac{1}{\sqrt{|\widetilde{\boldsymbol{\Sigma}}_w|}} \exp\left(-\frac{1}{2}(\mathbf{x}_w - \widetilde{\mathbf{h}}_w)^{\mathrm{T}} \widetilde{\boldsymbol{\Sigma}}_w^{-1} (\mathbf{x}_w - \widetilde{\mathbf{h}}_w)\right), \tag{33}$$

with

$$\widetilde{\mathbf{h}}_w = \mathbf{h}_w\left(\tilde{\boldsymbol{\theta}}_{\max}(\mathbf{w}), \mathbf{w}\right)$$
$$\widetilde{\boldsymbol{\Sigma}}_w = \boldsymbol{\Sigma}_w\left(\tilde{\boldsymbol{\theta}}_{\max}(\mathbf{w}), \mathbf{w}\right). \tag{34}$$

Let k and l denote, respectively, the indices of the k'-th and the l'-th 1 in $\mathbf{w}$. Then the k'-th element of $\widetilde{\mathbf{h}}_w$ is obtained as $\|\tilde{\boldsymbol{\theta}}_{\max}(\mathbf{w}) - \mathbf{z}_k\|$, and the (k', l')-th element of $\widetilde{\boldsymbol{\Sigma}}_w$ is obtained as $\rho_{kl}\widetilde{\sigma}_k\widetilde{\sigma}_l$, with

$$\widetilde{\sigma}_m^2 = \sigma_0^2\left(\left\|\tilde{\boldsymbol{\theta}}_{\max}(\mathbf{w}) - \mathbf{z}_m\right\| + \zeta\right)^{\gamma}. \tag{35}$$

For finding a simple and effective proposal distribution $p_{\mathrm{rem}}(m|\mathbf{w}^{(j-1)})$ (cf. the discussion above (23)), we first simplify (33) by neglecting all the off-diagonal elements of $\widetilde{\boldsymbol{\Sigma}}_w$, which leads to

$$p_{\mathrm{diag}}(\mathbf{w}) \propto \left(\prod_{m=1}^{D} \frac{1}{\widetilde{\sigma}_m^{w_m}}\right) \exp\left(-\frac{1}{2}\sum_{m=1}^{D} \frac{w_m(x_m - \widetilde{h}_m)^2}{\widetilde{\sigma}_m^2}\right)$$
$$= \prod_{m=1}^{D} \underbrace{\exp\left(-\frac{w_m}{2}\left(\frac{(x_m - \widetilde{h}_m)^2}{\widetilde{\sigma}_m^2} + \log\left(\widetilde{\sigma}_m^2\right)\right)\right)}_{p_m(w_m)}.$$

Note that $p_m(0) = 1$ for all m. Following (23), we design the probabilities $p_{\mathrm{rem}}(m|\mathbf{w}^{(j-1)})$ of the indices m based on $1/p_m(1)$:

$$p_{\mathrm{rem}}\left(m \,\middle|\, \mathbf{w}^{(j-1)}\right) \propto \log\left(\frac{1}{p_m(1)}\right) - p_{\min} \tag{36}$$
$$\propto \frac{1}{2}\left(\frac{(x_m - \widetilde{h}_m)^2}{\widetilde{\sigma}_m^2} + \log\left(\widetilde{\sigma}_m^2\right)\right) - p_{\min}, \tag{37}$$

for all $m \in \{1, 2, \ldots, D\}$ such that $w_m^{(j-1)} = 1$. The logarithm in (36) is used in order to make the distribution flatter, thus allowing more variation and improving the results, according to our simulations. To ensure that all probabilities are non-negative, we subtract a constant $p_{\min}$ in (36) such that the smallest value of $p_{\mathrm{rem}}(m|\mathbf{w}^{(j-1)})$ equals zero. In (37), $\widetilde{h}_m$ and $\widetilde{\sigma}_m^2$ are obtained by inserting $\mathbf{w}^{(j-1)}$ for $\mathbf{w}$ in (34) and (35), respectively. The distribution $p_{\mathrm{rem}}(m|\mathbf{w}^{(j-1)})$ is normalized such that its sum over the indices m where $w_m^{(j-1)} = 1$ equals 1.

The proposed MH algorithm for robust censoring in target localization follows Algorithm 1, using (37) for generating $m^{(\mathrm{rem})}$ and inserting (33) and (37) into (25) and (26) for calculating α_j.

C. Numerical Results

Reference Method: AD: We compare the performance of our method to that of the straightforward AD approach presented at the beginning of Section III. Due to (30), calculating (12) for target localization specializes to

$$\mathbf{w}[i] = \underset{\mathbf{w}\in\mathcal{W}}{\arg\max}\, \mathcal{N}\left(\mathbf{x}_w; \mathbf{h}_w\left(\boldsymbol{\theta}[i-1], \mathbf{w}\right), \boldsymbol{\Sigma}_w\left(\boldsymbol{\theta}[i-1], \mathbf{w}\right)\right). \tag{38}$$

In order to calculate this approximately, we use the same simplification as described above (37), i.e., we neglect the off-diagonal elements of $\boldsymbol{\Sigma}_w(\boldsymbol{\theta}[i-1], \mathbf{w})$. Denoting the m-th diagonal element of $\boldsymbol{\Sigma}(\boldsymbol{\theta}[i-1])$ by

$$\sigma_m^2[i-1] = \sigma_0^2\left(\|\boldsymbol{\theta}[i-1] - \mathbf{z}_m\| + \zeta\right)^{\gamma},$$

we obtain

$$\mathbf{w}[i] = \underset{\mathbf{w}\in\mathcal{W}}{\arg\max} \prod_{m=1}^{D} \left(\mathcal{N}\left(x_m; h_m\left(\boldsymbol{\theta}[i-1]\right), \sigma_m^2[i-1]\right)\right)^{w_m}$$
$$= \underset{\mathbf{w}\in\mathcal{W}}{\arg\min} \sum_{m=1}^{D} \frac{w_m}{2} \times \underbrace{\left(\frac{\left(x_m - h_m\left(\boldsymbol{\theta}[i-1]\right)\right)^2}{\sigma_m^2[i-1]} + \log\left(\sigma_m^2[i-1]\right)\right)}_{\xi_m}. \tag{39}$$

As is easily verified, (39) amounts to finding the d smallest values among $\xi_1, \xi_2, \ldots, \xi_D$ and setting the corresponding elements of $\mathbf{w}[i]$ to 1 and the remaining elements of $\mathbf{w}[i]$ to 0. For solving (11) for target localization, we use the same steps as for solving (32). In our implementation, the AD method is significantly less computationally complex than the proposed MH method.

Reference Method: SCOR: As a second reference method besides AD, we extended SCOR to the non-linear problem of target localization. To this end, we replace (31) for this method with

$$(\hat{\boldsymbol{\theta}}, \hat{\mathbf{o}}) = \underset{(\boldsymbol{\theta},\mathbf{o})\in\mathbb{R}^N\times\mathbb{R}^D}{\arg\min} \left\|\left(\boldsymbol{\Sigma}(\boldsymbol{\theta})\right)^{-1/2}\left(\mathbf{x} - \mathbf{h}(\boldsymbol{\theta}) - \mathbf{o}\right)\right\|^2 + \log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \lambda\|\mathbf{o}\|_1. \tag{40}$$

Direct application of SCOR to (40) is not possible because $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$ and because $\mathbf{h}(\boldsymbol{\theta})$ is a non-linear function of $\boldsymbol{\theta}$. Instead, we resort to an approach analogous to (11), (12); we alternately minimize the multivariate cost function in (40) with respect to either $\boldsymbol{\theta}$ or $\mathbf{o}$ while the respective other parameter is fixed, thus converging to a local minimum of the cost function. More specifically, in iteration i, we calculate

$$\mathbf{o}[i] = \underset{\mathbf{o}\in\mathbb{R}^D}{\arg\min} \left\|\left(\boldsymbol{\Sigma}\left(\boldsymbol{\theta}[i-1]\right)\right)^{-1/2}\left(\mathbf{x} - \mathbf{h}\left(\boldsymbol{\theta}[i-1]\right) - \mathbf{o}\right)\right\|^2 + \lambda\|\mathbf{o}\|_1 \tag{41}$$

$$\boldsymbol{\theta}[i] = \underset{\boldsymbol{\theta}\in\mathbb{R}^N}{\arg\min} \left\| \left(\boldsymbol{\Sigma}(\boldsymbol{\theta})\right)^{-1/2} \left(\mathbf{x} - \mathbf{h}(\boldsymbol{\theta}) - \mathbf{o}[i]\right) \right\|^2 + \log\left|\boldsymbol{\Sigma}(\boldsymbol{\theta})\right|. \quad (42)$$

Here, $\mathbf{o}[i]$ can be calculated using LASSO [47] like in the original SCOR for linear problems. In each iteration, we choose a suitable value of λ. Guidelines for a robust way of choosing λ are given in [13]. However, since this SCOR method was already significantly more complex than the proposed MH method, we do not follow [13] here but instead choose λ heuristically. In each iteration, we set λ equal to the largest absolute value of the vector $0.6\ \mathrm{diag}(\boldsymbol{\sigma}_n)(\mathbf{x} - \mathbf{h}(\boldsymbol{\theta}[i-1]))$, where $\boldsymbol{\sigma}_n$ contains the diagonal elements of $(\boldsymbol{\Sigma}(\boldsymbol{\theta}[i-1]))^{-1/2}$. As is easily verified by comparing (42) and (32), $\boldsymbol{\theta}[i]$ is obtained by calculating $\boldsymbol{\theta}_{\max}(\mathbf{1})$ with $\mathbf{x}$ replaced by $\mathbf{x} - \mathbf{o}[i]$. The algorithm is initialized with $\boldsymbol{\theta}[0] = \boldsymbol{\theta}_{\max}(\mathbf{w}_{\text{init}})$, where $\mathbf{w}_{\text{init}}$ is randomly drawn from a uniform distribution over $\mathcal{W}$. The algorithm terminates after iteration i if $\mathbf{w}[i] = \mathbf{w}[i-1]$, where $\mathbf{w}[i]$ is calculated from $\mathbf{o}[i]$ by finding the d largest absolute values among the elements of $\mathbf{o}[i]$ and setting the corresponding elements of $\mathbf{w}[i]$ to 1 and the remaining elements of $\mathbf{w}[i]$ to 0. After its final iteration i, the algorithm returns $\hat{\boldsymbol{\theta}}_{\text{SCOR}} = \boldsymbol{\theta}[i]$ and $\hat{\mathbf{w}}_{\text{SCOR}} = \mathbf{w}[i]$. As mentioned above, in our implementations, the computational complexity of this method is significantly higher than that of the proposed MH method. The reason is that the calculations in the MH method are based on vectors of length d or matrices of size $d \times d$, e.g., $\mathbf{x}_w$, $\widetilde{\mathbf{h}}_w$, or $\widetilde{\boldsymbol{\Sigma}}_w$, whereas the calculations in the SCOR method are based on vectors of length D or matrices of size $D \times D$, e.g., $\mathbf{x}$, $\mathbf{h}(\boldsymbol{\theta})$, or $\boldsymbol{\Sigma}(\boldsymbol{\theta})$.

Simulation Setup: To assess and compare the performance of the methods described above, we generated several hundred measurement vectors according to (27), using $D = 1000$ and $N = 2$. In the following, all lengths are normalized with respect to a unit length of 1m. For each measurement vector, the true location $\boldsymbol{\theta}$ and the sensor locations $\mathbf{z}_1, \ldots, \mathbf{z}_D$ were individually generated from a uniform distribution on $[0, 100]^2$. Each noise vector $\mathbf{n}$ was generated using $\zeta = 5$ and $\gamma = 1$. The correlation coefficients ρ_{kl} were generated individually for each noise vector. To this end, we first generated a matrix $\mathbf{R}$ from a uniform distribution on $[0, \sqrt{4\bar{\rho}/D}]^{D\times D}$, where $\bar{\rho}$ was varied in different experiments. Then, the coefficient ρ_{kl} for $k \neq l$ was obtained as the (k,l)-th element of $\mathbf{R}^{\mathrm{T}}\mathbf{R}$. The resulting coefficients ρ_{kl} for $k \neq l$ were distributed closely around their mean $\bar{\rho}$, while we set $\rho_{kk} = 1$ for $k = 1, 2, \ldots, D$. In each data vector, o out of the $D = 1000$ measurements were contaminated with outliers, by adding zero-mean Gaussian noise with variance σ^2_{out}. In different experiments, d, o, σ^2_{out} and $\bar{\rho}$ were varied to study the behavior of the methods in different settings.

For each data vector, $\boldsymbol{\theta}$ and $\mathbf{w}$ were calculated according to the proposed method, i.e., Algorithm 1, with $J = 200$ iterations and $\beta = 0.005$. The corresponding curves are labeled as MH. We compare the proposed method with the two reference methods described above, labeled as SCOR and as AD, respectively. The performance measures we assess are the empirical root-mean-square error (RMSE) of $\hat{\boldsymbol{\theta}}$ obtained by averaging over 200 experiments, and the average number $\bar{o}_w$ of outlier measurements among the selected measurements $\mathbf{x}_w$ according to $\hat{\mathbf{w}}$, i.e., outliers that were not successfully rejected.

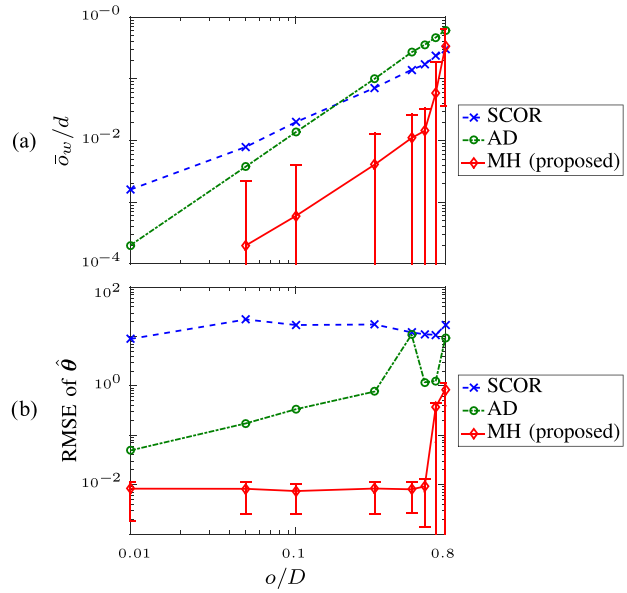

Fig. 2. Performance for different degrees of outlier contamination o/D: (a) Average rate of outliers $\bar{o}_w/d$ among the selected measurements $\mathbf{x}_w$ corresponding to $\hat{\mathbf{w}}$, (b) Empirical RMSE of $\hat{\boldsymbol{\theta}}$. For the MH method, vertical bars mark an interval as wide as the empirical standard deviation on each side of the respective average. For $o/D = 0.01$, the MH method yields $\bar{o}_w/d = 0$, i.e., the selected measurements were outlier-free in all 200 experiments.

Normalizing $\bar{o}_w$ with d yields the average residual rate of outlier contamination among the selected measurements, which can be compared to the unprocessed rate of outlier contamination o/D.

Simulation Results: Fig. 2 shows the average residual outlier contamination rate $\bar{o}_w/d$ as well as the RMSE of $\hat{\boldsymbol{\theta}}$ for different values of o, i.e., different degrees of outlier contamination o/D between 1% and 80%. Here, we used $d = 50$, $\sigma^2_{\text{out}} = 10$, $\sigma^2_0 = 10^{-5}$, and $\bar{\rho} = 0.2$. We can see that, in terms of $\bar{o}_w/d$, all methods perform well at lower contamination o/D, successfully eliminating the outliers almost completely. The proposed method is clearly the most robust, showing the smallest $\bar{o}_w/d$, up to very high outlier rates above 70%. In terms of $\hat{\boldsymbol{\theta}}$, the proposed method outperforms the other methods significantly for most degrees of outlier contamination o/D. The performance gap only becomes smaller around $o/D = 70\%$ and above. Interestingly, the error of $\hat{\boldsymbol{\theta}}$ does not appear to depend on o/D in the proposed method (up to about $o/D = 60\%$) and SCOR (within the range studied here). In this sense both methods are robust to higher outlier contamination, but at very different error levels.

Fig. 3 shows results from 200 experiments where o/D is fixed at 10% and d is varied between 10 and 150, leading to different compression rates d/D. As in Fig. 2, the proposed method clearly outperforms the other methods in terms of both $\bar{o}_w/d$ (at least for smaller d) and $\hat{\boldsymbol{\theta}}$ (for the entire range of d). In the proposed method and in SCOR, the performance shows very weak or no visible dependence on d, while the performance of AD degrades when fewer measurements are selected. For d larger than about 100, AD performs similarly to the proposed method in terms of $\bar{o}_w/d$, but it still clearly yields a larger estimation error in terms of $\hat{\boldsymbol{\theta}}$.

In Fig. 4, we show the dependence of the RMSE of $\hat{\boldsymbol{\theta}}$ and of the average residual outlier contamination rate $\bar{o}_w/d$ on the outlier strength σ^2_{out}, using $d = 50$, $o/D = 10\%$, $\sigma^2_0 = 10^{-5}$, and $\bar{\rho} = 0.2$. We can see that AD yields a roughly constant $\bar{o}_w/d$

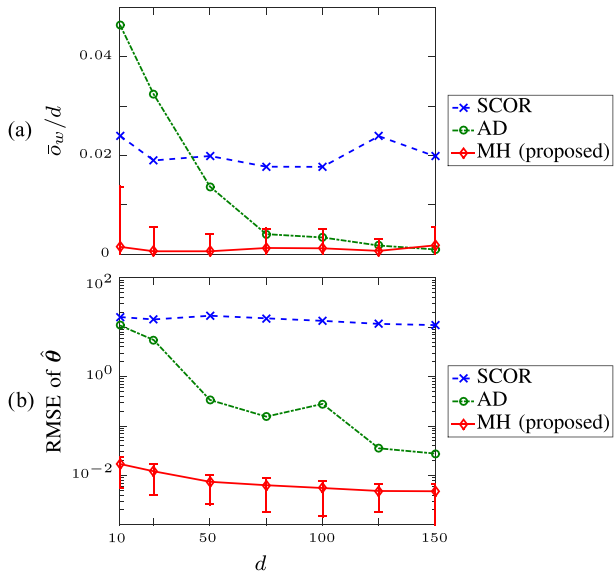

Fig. 3. Performance for different numbers d of selected measurements out of $D = 1000$: (a) Average rate of outliers $\bar{o}_w/d$ among the selected measurements $\mathbf{x}_w$ corresponding to $\hat{\mathbf{w}}$, (b) Empirical RMSE of $\hat{\boldsymbol{\theta}}$. For the MH method, vertical bars mark an interval as wide as the empirical standard deviation on each side of the respective average.

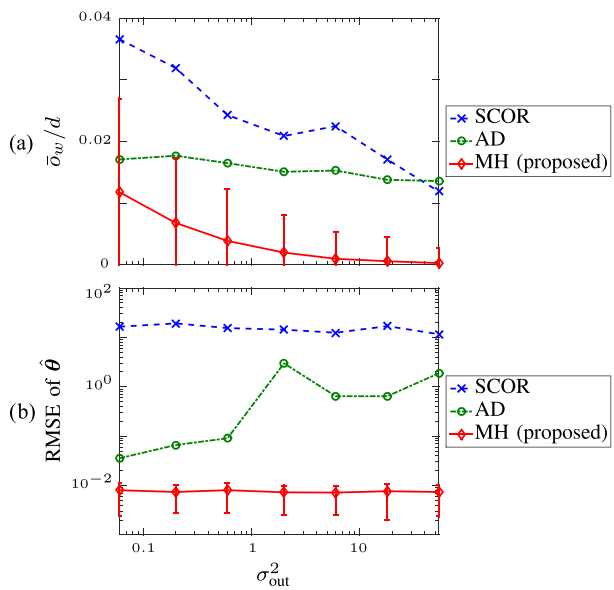

Fig. 4. Performance for different outlier variances σ^2_{out}: (a) Average rate of outliers $\bar{o}_w/d$ among the selected measurements $\mathbf{x}_w$ corresponding to $\hat{\mathbf{w}}$, (b) Empirical RMSE of $\hat{\boldsymbol{\theta}}$. For the MH method, vertical bars mark an interval as wide as the empirical standard deviation on each side of the respective average.

for different σ^2_{out}, while its estimate $\hat{\boldsymbol{\theta}}$ becomes worse for larger σ^2_{out}. Both SCOR and the proposed method, on the other hand, improve in terms of $\bar{o}_w/d$ as σ^2_{out} increases, while their estimation errors of $\hat{\boldsymbol{\theta}}$ appear fairly independent of σ^2_{out}. This result seems intuitive, since smaller outliers can more easily be missed on the one hand, but they do not cause as much degradation in the observation on the other hand. In both performance measures, the proposed method performs consistently much better than SCOR, and consistently better than AD.

Fig. 5 shows the average residual outlier contamination rate $\bar{o}_w/d$ as well as the RMSE of $\hat{\boldsymbol{\theta}}$ for different values of σ^2_0, i.e., different noise levels. Here, we used $d = 50$, $o/D = 10\%$, $\sigma^2_{\text{out}} = 10^4$, and $\bar{\rho} = 0.2$. We can see that, in terms of $\bar{o}_w/d$,

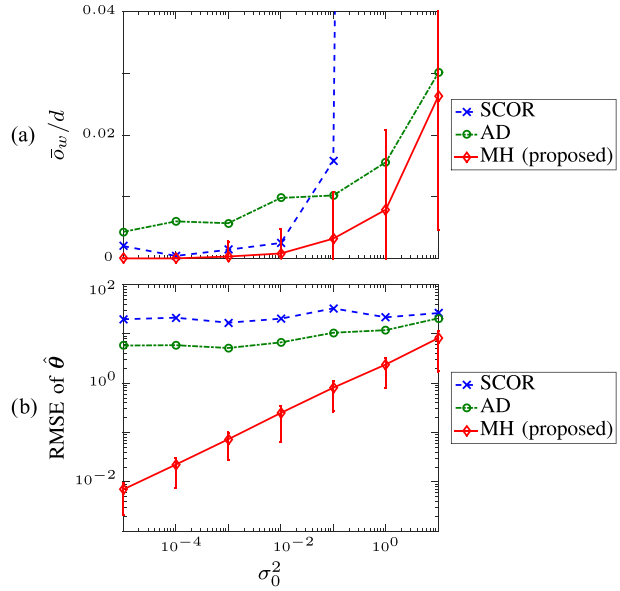

Fig. 5. Performance for different noise levels σ^2_0: (a) Average rate of outliers $\bar{o}_w/d$ among the selected measurements $\mathbf{x}_w$ corresponding to $\hat{\mathbf{w}}$, (b) Empirical RMSE of $\hat{\boldsymbol{\theta}}$. For the MH method, vertical bars mark an interval as wide as the empirical standard deviation on each side of the respective average. For $\sigma^2_0 = 10^0$ and $\sigma^2_0 = 10^1$, the SCOR method yields, respectively, $\bar{o}_w/d = 0.82$ and $\bar{o}_w/d = 1$.

all methods produce rather flat curves at lower noise levels and significant increases at higher noise levels. The increase is most drastic in the SCOR method, where it reaches $\bar{o}_w/d = 100\%$, i.e., all selected observations are outlier-contaminated. Interestingly, this drastic failure at outlier rejection does not correspond to a significant increase of the RMSE of $\hat{\boldsymbol{\theta}}$ compared to lower noise levels. The proposed MH method consistently performs best in both performance measures. In terms of $\hat{\boldsymbol{\theta}}$, the performance gap is largest at low noise levels and steadily decreases with increasing noise.

Fig. 6 studies the effect of the correlations among different measurements. We vary $\bar{\rho}$ between 0 and 0.6, using $d = 50$, $o/D = 10\%$, $\sigma^2_{\text{out}} = 10$, and $\sigma^2_0 = 10^{-5}$. We can see that the proposed method handles correlations best, achieving smaller errors and stronger outlier rejection than the reference methods at all levels of correlation. While the average residual outlier contamination rate $\bar{o}_w/d$ of the proposed method appears fairly invariant to $\bar{\rho}$ (by contrast to the reference methods), the error of $\hat{\boldsymbol{\theta}}$ increases slightly at higher correlation levels.

Finally, in Fig. 7 we study the behavior of the proposed MH method over the number of iterations J. The results are averaged from 10 000 experiments using $d = 50$, $o/D = 10\%$, $\sigma^2_{\text{out}} = 10$, $\sigma^2_0 = 10^{-5}$, and $\bar{\rho} = 0.2$. Fig. 7(a) shows how many of the $D = 1000$ observations are used within the first J iterations. Fig. 7(b) shows the dependence of the RMSE of $\hat{\boldsymbol{\theta}}$ and of the average residual outlier contamination rate $\bar{o}_w/d$ on J. We can see that the number of observations converges towards $D = 1000$ steadily but rather slowly. By contrast, the error decreases very quickly in the first $\sim$50 iterations and then improves only marginally in further iterations. We can conclude from the slow decrease of the errors at later iterations that a prohibitively large number of iterations may be required to guarantee a high probability of obtaining the unique optimal $\hat{\mathbf{w}}$ among the $\binom{D}{d} \approx 10^{85}$ elements of $\mathcal{W}$. On the other hand, the

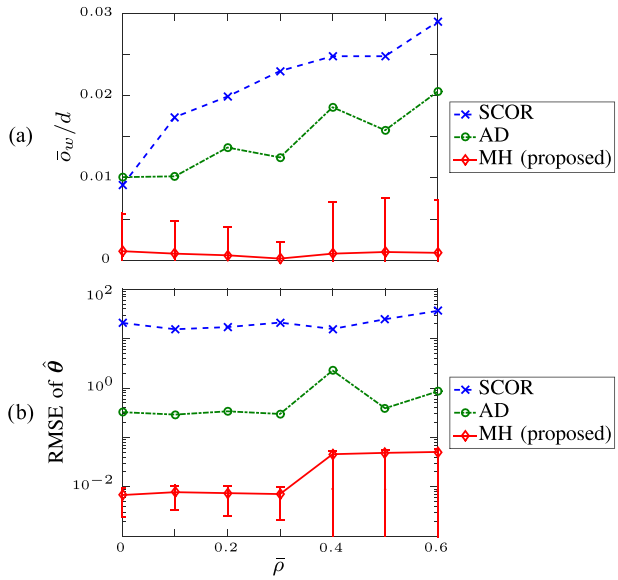

Fig. 6. Performance for different values of average measurement correlation $\bar{\rho}$: (a) Average rate of outliers $\bar{o}_w/d$ among the selected measurements $\mathbf{x}_w$ corresponding to $\hat{\mathbf{w}}$, (b) Empirical RMSE of $\hat{\boldsymbol{\theta}}$. For the MH method, vertical bars mark an interval as wide as the empirical standard deviation on each side of the respective average.

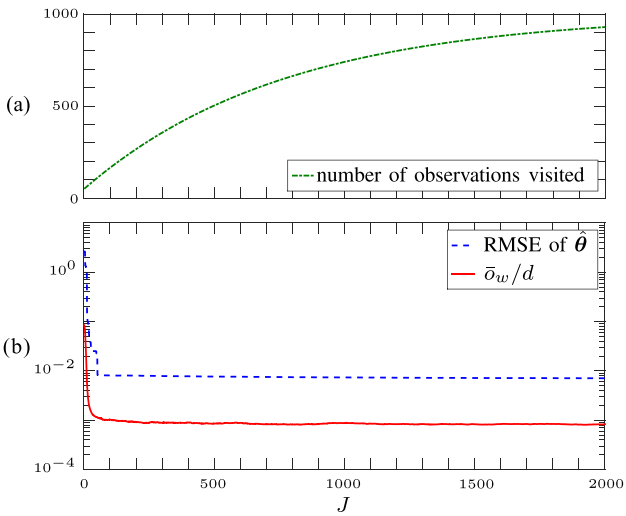

Fig. 7. Behavior of the proposed MH method over the number of iterations J: (a) Number of observations x_m that are used within the first J iterations, (b) Empirical RMSE of $\hat{\boldsymbol{\theta}}$ and average rate of outliers $\bar{o}_w/d$ among the selected measurements $\mathbf{x}_w$ after J iterations.

outlier rate is reduced from $o/D = 10\%$ to $\bar{o}_w/d = 0.1\%$ within only 122 iterations, using only 188.8 of the 1000 observations.

The average computation times for an unoptimized MATLAB R2014b 64-bit implementation on a 2.5-GHz Intel Core i5 processor were 0.41s for the proposed MH method (with $J = 200$), 0.03s for AM, and 30.22s for SCOR.

V. Hybrid Model-Data-Driven Scheme

The problem formulation for robust censoring presented in (3) links the applications of outlier rejection and data-driven dimensionality reduction, yielding a solution that is optimal in the maximum likelihood sense. Similarly as in other methods for outlier rejection or for data-driven dimensionality reduction, the decision criterion in (3) does not take into account the resulting inference performance in terms of the mean squared error (MSE). On the other hand, dimensionality reduction schemes that are optimal in terms of the MSE, i.e., model-driven schemes, are not robust to outliers. In this section, we propose to extend the decision criterion in (3) such that it reduces the resulting MSE. We present the corresponding modifications of the proposed MH method for robust censoring, and we give an example where the extended *hybrid model-data-driven* sensing scheme indeed leads to improved performance.

For ease of exposition, we consider a linear inverse problem. Recall that this model was already introduced in (4)–(6). As stated there, the robust censoring problem formulated in (3) simplifies for the linear inverse problem (4) to

$$(\hat{\boldsymbol{\theta}}, \hat{\mathbf{w}}) = \underset{(\boldsymbol{\theta},\mathbf{w})\in\mathbb{R}^N\times\mathcal{W}}{\arg\min} \|\mathbf{x}_w - \mathbf{A}_w\boldsymbol{\theta}\|^2.$$

In analogy to (14), (15), we can write this as

$$\begin{aligned}\hat{\mathbf{w}} &= \underset{\mathbf{w}\in\mathcal{W}}{\arg\min} \underset{\boldsymbol{\theta}\in\mathbb{R}^N}{\min} \|\mathbf{x}_w - \mathbf{A}_w\boldsymbol{\theta}\|^2 \\ &= \underset{\mathbf{w}\in\mathcal{W}}{\arg\min} \|\mathbf{x}_w - \mathbf{A}_w\boldsymbol{\theta}_{\max}(\mathbf{w})\|^2 \qquad (43)\\ \hat{\boldsymbol{\theta}} &= \boldsymbol{\theta}_{\max}(\hat{\mathbf{w}}), \qquad (44)\end{aligned}$$

with (cf. (13))

$$\begin{aligned}\boldsymbol{\theta}_{\max}(\mathbf{w}) &= \underset{\boldsymbol{\theta}\in\mathbb{R}^N}{\arg\min} \|\mathbf{x}_w - \mathbf{A}_w\boldsymbol{\theta}\|^2 \\ &= \left(\mathbf{A}_w{}^{\mathrm{T}}\mathbf{A}_w\right)^{-1}\mathbf{A}_w{}^{\mathrm{T}}\mathbf{x}_w. \qquad (45)\end{aligned}$$

Inserting (45) into (43) yields

$$\hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathcal{W}}{\arg\min}\ \tilde{r}(\mathbf{w}), \qquad (46)$$

with

$$\tilde{r}(\mathbf{w}) = \mathbf{x}_w^{\mathrm{T}}\left(\mathbf{I} - \mathbf{A}_w\left(\mathbf{A}_w{}^{\mathrm{T}}\mathbf{A}_w\right)^{-1}\mathbf{A}_w{}^{\mathrm{T}}\right)\mathbf{x}_w. \qquad (47)$$

We will now modify (47) such that it also takes into account the MSE of $\hat{\boldsymbol{\theta}}$, while (44) and (45) remain unchanged. Due to (44), we can express the MSE of $\hat{\boldsymbol{\theta}}$ as a function of $\hat{\mathbf{w}}$:

$$\mathrm{MSE}\{\hat{\boldsymbol{\theta}};\hat{\mathbf{w}}\} = \mathbb{E}\left\{\|\boldsymbol{\theta}_{\max}(\hat{\mathbf{w}}) - \boldsymbol{\theta}\|^2\right\}.$$

Using (45) and (4), this can be shown to yield

$$\begin{aligned}\mathrm{MSE}\{\hat{\boldsymbol{\theta}};\hat{\mathbf{w}}\} &= \sigma^2\ \mathrm{tr}\left\{\left(\mathbf{A}_w{}^{\mathrm{T}}\mathbf{A}_w\right)^{-1}\right\} \\ &= \sigma^2\ \mathrm{tr}\left\{\left(\sum_{m=1}^{D}\hat{w}_m\mathbf{a}_m\mathbf{a}_m^{\mathrm{T}}\right)^{-1}\right\}.\end{aligned}$$

By adding

$$f(\mathbf{w}) = \frac{\mathrm{MSE}\{\hat{\boldsymbol{\theta}};\mathbf{w}\}}{\sigma^2} = \mathrm{tr}\left\{\left(\sum_{m=1}^{D} w_m\mathbf{a}_m\mathbf{a}_m^{\mathrm{T}}\right)^{-1}\right\} \qquad (48)$$

as a penalty term in the cost function in (47), we obtain a hybrid model-data-driven sensing scheme which jointly minimizes the negative likelihood function and the MSE of $\hat{\boldsymbol{\theta}}$:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathcal{W}}{\arg\min}\ \tilde{r}(\mathbf{w}) + \lambda f(\mathbf{w}). \qquad (49)$$

Here, λ is a tuning parameter. In particular, $\lambda \to 0(\infty)$ results in the related data (model)-driven scheme. The above optimization problem is non-convex in $\mathbf{w}$. A generalization of the hybrid model-data-driven sensing scheme is discussed in the following remark.

Remark (Non-Linear Model): The proposed hybrid model-data-driven scheme can be generalized to more complicated observation models, e.g., non-Gaussian and/or non-additive noise models. In a more general hybrid scheme, we will replace $\tilde{r}(\mathbf{w})$ in (49) with a negative log-likelihood function and $f(\mathbf{w})$ with the Cramér-Rao bound as in [6], and the minimization will be over both $\boldsymbol{\theta}$ and $\mathbf{w} \in \mathcal{W}$.

Before we present the proposed MH method for solving (49), we discuss an alternative method based on convex relaxation, which will serve as a performance benchmark.

A. Convex Relaxation

The optimization problem (49) can be equivalently written in the epigraph form as

$$\begin{aligned} \underset{\mathbf{w}\in\mathcal{W},t_1,t_2}{\arg\min} \quad & t_1 + \lambda t_2 && \text{(50a)} \\ \text{subject to} \quad & \tilde{r}(\mathbf{w}) \le t_1, && \text{(50b)} \\ & f(\mathbf{w}) \le t_2, && \text{(50c)} \end{aligned}$$

with auxiliary variables $t_1 \in \mathbb{R}$ and $t_2 \in \mathbb{R}$, where the constraint (50c) is convex in $\mathbf{w}$. We relax the non-convex constraint set $\mathcal{W}$ to its best convex approximation

$$\mathcal{W}_c = \{\mathbf{w} \mid \mathbf{1}^{\mathrm{T}}\mathbf{w} = d, 0 \le w_m \le 1, m = 1, 2, \ldots, D\}.$$

Using the Schur complement and the property that $\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} = \mathrm{diag}(\mathbf{w})$, the constraint (50b) can be equivalently expressed as

$$\begin{bmatrix} \mathbf{A}^{\mathrm{T}}\mathrm{diag}(\mathbf{w})\mathbf{A} & \mathbf{A}^{\mathrm{T}}\mathrm{diag}(\mathbf{w})\mathbf{x} \\ \mathbf{x}^{\mathrm{T}}\mathrm{diag}(\mathbf{w})\mathbf{A} & t_1 - \mathbf{x}^{\mathrm{T}}\mathrm{diag}(\mathbf{w})\mathbf{x} \end{bmatrix} \succeq \mathbf{0},$$

which is convex and linear in $\mathbf{w}$ and t_1. The convex relaxed hybrid model-data-driven sensing problem then becomes

$$\begin{aligned} \underset{\mathbf{w}\in\mathcal{W}_c,t_1,t_2}{\arg\min} \quad & t_1 + \lambda t_2 \\ \text{subject to} \quad & \begin{bmatrix} \mathbf{A}^{\mathrm{T}}\mathrm{diag}(\mathbf{w})\mathbf{A} & \mathbf{A}^{\mathrm{T}}\mathrm{diag}(\mathbf{w})\mathbf{x} \\ \mathbf{x}^{\mathrm{T}}\mathrm{diag}(\mathbf{w})\mathbf{A} & t_1 - \mathbf{x}^{\mathrm{T}}\mathrm{diag}(\mathbf{w})\mathbf{x} \end{bmatrix} \succeq \mathbf{0}, \\ & f(\mathbf{w}) \le t_2. \end{aligned} \tag{51}$$

The solution of this problem is not Boolean; however, an approximate Boolean solution can be obtained by using deterministic or randomized rounding as discussed in [6].

Contrary to the solver presented in this subsection, the proposed MH method for solving (49), which will be introduced next, does not use any convex relaxations.

B. MH Sampler

For the data-driven MH sampler as proposed in Section III, the target distribution (cf. (16), (5)) for the linear problem considered in this section becomes

$$\begin{aligned} p_{\mathrm{t}}(\mathbf{w}) &\propto \exp\left(-\frac{1}{2\sigma^2}\left\|\mathbf{x}_w - A_w\boldsymbol{\theta}_{\max}(\mathbf{w})\right\|^2\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\,\tilde{r}(\mathbf{w})\right), \end{aligned}$$

where we used (45) and (47). In analogy to the hybrid problem formulation in (49), we can modify the above $p_{\mathrm{t}}(\mathbf{w})$ to

$$p_{\mathrm{t,hybrid}}(\mathbf{w}) \propto \exp\left(-\frac{1}{2\sigma^2}\left(\tilde{r}(\mathbf{w}) + \lambda f(\mathbf{w})\right)\right), \tag{52}$$

in order to design a MH sampler for hybrid model-data-driven robust censoring.

Analogously to Section IV-B, we design a simple and effective proposal distribution $p_{\mathrm{rem}}(m|\mathbf{w}^{(j-1)})$ (cf. the discussion above (23)) by simplifying the target distribution to a product of factors $p_m(w_m)$ such that each factor depends only on one element of $\mathbf{w}$. To this end, we first replace $f(\mathbf{w})$ from (48) with

$$\tilde{f}(\mathbf{w}) = \sum_{m=1}^{D} w_m \left(\mathrm{tr}\left\{\mathbf{a}_m\mathbf{a}_m^{\mathrm{T}}\right\}\right)^{-1} = \sum_{m=1}^{D} w_m \frac{1}{\|\mathbf{a}_m\|^2}.$$

Furthermore replacing $\boldsymbol{\theta}_{\max}(\mathbf{w})$ with a fixed $\tilde{\boldsymbol{\theta}}$ that does not depend on $\mathbf{w}$—we use $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}_{\max}(\mathbf{w}^{(j-1)})$—leads to

$$\begin{aligned} p_{\mathrm{diag}}(\mathbf{w}) &\propto \exp\left(-\frac{\left\|\mathbf{x}_w - \mathbf{A}_w\tilde{\boldsymbol{\theta}}\right\|^2 + \lambda\tilde{f}(\mathbf{w})}{2\sigma^2}\right) \\ &= \prod_{m=1}^{D} \underbrace{\exp\left(-\frac{w_m}{2\sigma^2}\left(\left(x_m - \mathbf{a}_m^{\mathrm{T}}\tilde{\boldsymbol{\theta}}\right)^2 + \lambda\frac{1}{\|\mathbf{a}_m\|^2}\right)\right)}_{p_m(w_m)}. \end{aligned}$$

Note that $p_m(0) = 1$ for all m. Following the discussion above (23), we set (analogously to (36)):

$$\begin{aligned} p_{\mathrm{rem}}\left(m \,\middle|\, \mathbf{w}^{(j-1)}\right) &\propto \log\left(\frac{1}{p_m(1)}\right) && \text{(53)} \\ &\propto \frac{1}{2\sigma^2}\left(\left(x_m - \mathbf{a}_m^{\mathrm{T}}\tilde{\boldsymbol{\theta}}\right)^2 + \lambda\frac{1}{\|\mathbf{a}_m\|^2}\right), && \text{(54)} \end{aligned}$$

for all $m \in \{1, 2, \ldots, D\}$ such that $w_m^{(j-1)} = 1$. Differently from (37), non-negativity of the probabilities is already guaranteed in (54) without adding a constant. The distribution $p_{\mathrm{rem}}(m|\mathbf{w}^{(j-1)})$ is normalized such that its sum over the indices m where $w_m^{(j-1)} = 1$ equals 1.

The proposed MH algorithm for hybrid model-data-driven robust censoring in linear problems follows Algorithm 1, using (54) for generating $m^{(\mathrm{rem})}$ and inserting (52) and (54) into (25) and (26) for calculating α_j.

The proposed MH approach is more flexible than the convex-relaxation-based method presented in Section V-A in that it can be extended to non-linear and/or non-Gaussian data models.

C. Numerical Results

In the numerical experiments presented in the following, we assess not only the two methods proposed above but also the hybrid censoring scheme itself, in comparison to the data-driven scheme and the model-driven scheme. To this end, we chose the dimensions of the problem small enough to allow for an exhaustive search over $\mathcal{W}$, thus obtaining the truly optimal estimate according to each sampling scheme, without potential inaccuracies due to some optimization method. We generated data according to (4) using $D = 16$, $N = 2$, $d = 4$, and $\sigma^2 = 10^{-2}$. For each experiment, the elements of $\boldsymbol{\theta}$ and $\mathbf{A}$ were generated

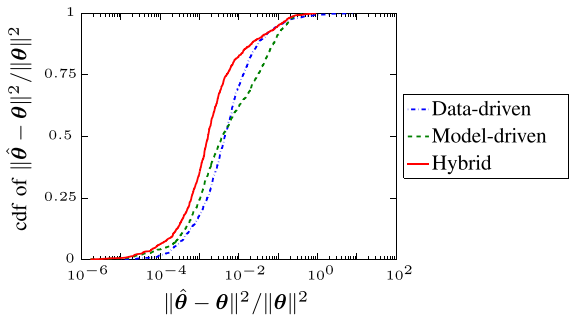

Fig. 8. Performance comparison of the three censoring schemes: cdf of $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2/\|\boldsymbol{\theta}\|^2$.

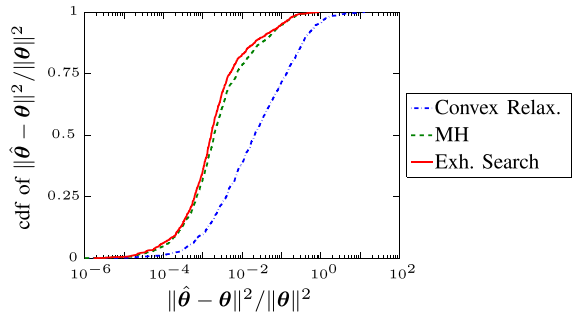

Fig. 9. Performance comparison for the two methods proposed above: cdf of $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2/\|\boldsymbol{\theta}\|^2$.

individually from zero-mean Gaussian distributions with variance 1. In each data vector, 3 out of the $D = 16$ measurements were contaminated with outliers, by adding zero-mean Gaussian noise with variance 1.

Evaluation of the Scheme: As mentioned in the previous subsection, both the data-driven and the model-driven scheme have their respective strengths and weaknesses. In a given scenario, either of them may perform better, depending on the strength and number of outlier measurements, among other parameters. Similarly, the proposed hybrid scheme may perform better or worse than either of the original schemes in a given scenario. It is interesting to note, however, that although the hybrid cost function is a linear combination of the two original cost functions, the best performance that can potentially be achieved by the hybrid scheme is not necessarily between the best performances achieved by the two original schemes. This is illustrated in Fig. 8, where we show the empirical cdf's of the error of the optimal estimate $\hat{\boldsymbol{\theta}}$ according to the three schemes. These optimal estimates were not obtained by applying the methods proposed above but by performing an exhaustive search over $\mathcal{W}$. For the hybrid scheme, we chose $\lambda = 5$. The cdf's were obtained from 1000 experiments. We can see that, here, the hybrid censoring scheme allows for a lower minimal error than both original schemes. Evidently, it exploits both the robustness with respect to outliers, which it shares with the data-driven scheme, and the information about the average reliability of each sensor, which it shares with the model-driven scheme.

Evaluation of the Methods: The two methods presented in Sections V-A and V-B are compared in Fig. 9. More specifically, we assess the empirical cdf's of their estimates $\hat{\boldsymbol{\theta}}$ from 1000 experiments using $\lambda = 5$. We can see that the error distribution achieved with the MH method almost coincides with that of exhaustive search, while the errors obtained by the convex relaxation method are larger.

VI. Conclusions

We proposed a novel joint approach to the two tasks of online data censoring and outlier-robust learning. This problem was formulated in terms of non-convex optimization, by jointly maximizing the likelihood of the reduced dataset with respect to both the inferred model parameters and the data selection vector. We showed that the specialization of our general approach to the linear Gaussian model is closely related to existing state-of-the-art methods for outlier rejection in this model. Based on the concept of Metropolis-Hastings sampling, we proposed a method for solving the general non-convex problem of robust censoring. We applied the proposed method to the problem of robust censoring for target localization and demonstrated its excellent performance in comparison to other approaches, as well as its high robustness with respect to the number and strength of outliers and other parameters. Finally, we also studied an extension to the original problem formulation, allowing us to improve the inference results in terms of average performance. We showed that the resulting hybrid censoring scheme may indeed perform better than both original schemes from which it was derived.

References

[1] G. Kail, S. P. Chepuri, and G. Leus, "Robust censoring for linear inverse problems," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Stockholm, Sweden, Jun.–Jul. 2015, pp. 495–499.
[2] P. J. Huber, *Robust Statist.*. New York, NY, USA: Wiley, 2009.
[3] J.-J. Fuchs, "An inverse problem approach to robust regression," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Phoenix, AZ, USA, Mar. 1999, pp. 1809–1812.
[4] A. Tajer, V. V. Veeravalli, and H. V. Poor, "Outlying sequence detection in large data sets: A data-driven approach," *IEEE Signal Process. Mag.*, vol. 31, no. 5, pp. 44–56, Sep. 2014.
[5] S. Joshi and S. Boyd, "Sensor selection via convex optimization," *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 451–462, Feb. 2009.
[6] S. P. Chepuri and G. Leus, "Sparsity-promoting sensor selection for non-linear measurement models," *IEEE Trans. Signal Process.*, vol. 63, no. 3, pp. 684–698, Feb. 2015.
[7] S. P. Chepuri and G. Leus, "Sparse sensing for distributed Gaussian detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, South Brisbane, Australia, Apr. 2015, pp. 2394–2398.
[8] S. P. Chepuri and G. Leus, "Sparsity-promoting adaptive sensor selection for non-linear filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Florence, Italy, May 2014, pp. 5100–5104.
[9] C. Rago, P. Willett, and Y. Bar-Shalom, "Censoring sensors: A low-communication-rate scheme for distributed detection," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 32, no. 2, pp. 554–568, Apr. 1996.
[10] E. J. Msechu and G. B. Giannakis, "Sensor-centric data reduction for estimation with WSNs via censoring and quantization," *IEEE Trans. Signal Process.*, vol. 60, no. 1, pp. 400–414, Jan. 2012.
[11] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and outlier Detection*. New York, NY, USA: Wiley, 2003.
[12] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981.
[13] G. B. Giannakis, G. Mateos, S. Farahmand, V. Kekatos, and H. Zhu, "USPACOR: Universal sparsity-controlling outlier rejection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Prague, Czech Republic, May 2011, pp. 1952–1955.
[14] V. Kekatos and G. B. Giannakis, "From sparse signals to sparse residuals for robust sensing," *IEEE Trans. Signal Process.*, vol. 59, no. 7, pp. 3355–3368, Jul. 2011.
[15] P. P. Vaidyanathan and P. Pal, "Sparse sensing with co-prime samplers and arrays," *IEEE Trans. Signal Process.*, vol. 59, no. 2, pp. 573–586, Feb. 2011.
[16] S. P. Chepuri and G. Leus, "Compression schemes for time-varying sparse signals," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Nov. 2014, pp. 1647–1651.

[17] R. G. Baraniuk, "Compressive sensing," *IEEE Signal Process. Mag.*, vol. 24, no. 4, pp. 118–121, Jul. 2007.
[18] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, Apr. 1970.
[19] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equations of state calculations by fast computing machines," *J. Chem. Phys.*, vol. 21, no. 6, pp. 1087–1092, Mar. 1953.
[20] S. Chib and E. Greenberg, "Understanding the Metropolis-Hastings algorithm," *Amer. Statist.*, vol. 49, no. 4, pp. 327–335, Nov. 1995.
[21] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. New York, NY, USA: Springer, 2004.
[22] W. R. Gilks, *Markov Chain Monte Carlo in Practice*. London, U.K.: Chapman & Hall, 1995.
[23] M. Davy, C. Doncarli, and J.-Y. Tourneret, "Supervised classification using MCMC methods," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Istanbul, Turkey, Jun. 2000, vol. 1, pp. I-33–I-36.
[24] S. Lesage, J.-Y. Tourneret, and P. M. Djuric, "Classification of digital modulations by MCMC sampling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Salt Lake City, UT, USA, May 2001, vol. 4, pp. IV-2553–IV-2556.
[25] J. Tang, Y. Zhang, J. Sun, J. Rao, W. Yu, Y. Chen, and A. C. M. Fong, "Quantitative study of individual emotional states in social networks," *IEEE Trans. Affective Comput.*, vol. 3, no. 2, pp. 132–144, Apr. 2012.
[26] T. Liang, G. J. Kacprzynski, K. Goebel, and G. Vachtsevanos, "Methodologies for uncertainty management in prognostics," in *Proc. IEEE Aerosp. Conf.*, Big Sky, MT, USA, Mar. 2009, pp. 1–12.
[27] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger, "On unbiased sampling for unstructured peer-to-peer networks," *IEEE/ACM Trans. Netw.*, vol. 17, no. 2, pp. 377–390, Apr. 2009.
[28] Z. Bar-Yossef and M. Gurevich, "Random sampling from a search engine's index," *J. ACM*, vol. 55, no. 5, pp. 91–164, Oct. 2008.
[29] V. Cevher, R. Velmurugan, and J. H. McClellan, "Acoustic multitarget tracking using direction-of-arrival batches," *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 2810–2825, Jun. 2007.
[30] V. Cevher and J. H. McClellan, "Fast initialization of particle filters using a modified Metropolis-Hastings algorithm: Mode-hungry approach," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Montreal, QC, Canada, May 2004, vol. 2, pp. II-129–II-132.
[31] A. C. Sankaranarayanan, A. Srivastava, and R. Chellappa, "Algorithmic and architectural optimizations for computationally efficient particle filtering," *IEEE Trans. Image Process.*, vol. 17, no. 5, pp. 737–748, May 2008.
[32] G. Kail, F. Hlawatsch, and C. Novak, "Efficient Bayesian detection of multiple events with a minimum-distance constraint," in *Proc. IEEE Statist. Signal Process. Workshop (SSP)*, Cardiff, Wales, U.K., Aug.–Sep. 2009, pp. 73–76.
[33] G. Kail, J.-Y. Tourneret, F. Hlawatsch, and N. Dobigeon, "Blind deconvolution of sparse pulse sequences under a minimum distance constraint: A partially collapsed Gibbs sampler method," *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 2727–2743, Jun. 2012.
[34] N. Dobigeon and J.-Y. Tourneret, "Bayesian orthogonal component analysis for sparse representation," *IEEE Trans. Signal Process.*, vol. 58, no. 5, pp. 2675–2685, May 2010.
[35] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc. Ser. B (Methodol.)*, vol. 39, no. 1, pp. 1–38, 1977.
[36] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA, USA: Morgan Kaufmann, 1988.
[37] A. Gelman and D. B. Rubin, "Inference from iterative simulation using multiple sequences," *Statist. Sci.*, vol. 7, no. 4, pp. 457–472, Nov. 1992.
[38] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. PAMI-6, no. 6, pp. 721–741, Nov. 1984.
[39] A. Gelfand and A. Smith, "Sampling-based approaches to calculating marginal densities," *J. Amer. Statist. Assoc.*, vol. 85, no. 410, pp. 398–409, Jun. 1990.
[40] D. A. van Dyk and T. Park, "Partially collapsed Gibbs samplers: Theory and methods," *J. Amer. Statist. Assoc.*, vol. 103, no. 482, pp. 790–796, Jun. 2008.
[41] T. Park and D. A. van Dyk, "Partially collapsed Gibbs samplers: Illustrations and applications," *J. Comput. Graph. Statist.*, vol. 18, no. 2, pp. 283–305, Jun. 2009.
[42] G. Kail, J.-Y. Tourneret, F. Hlawatsch, and N. Dobigeon, "A partially collapsed Gibbs sampler for parameters with local constraints," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, TX, USA, Mar. 2010, pp. 3886–3889.
[43] N. Patwari, J. N. Ash, S. Kyperountas, A. O. Hero, R. L. Moses, and N. S. Correal, "Locating the nodes: Cooperative localization in wireless sensor networks," *IEEE Signal Process. Mag.*, vol. 22, no. 4, pp. 54–69, Jul. 2005.
[44] F. Gustafsson and F. Gunnarsson, "Mobile positioning using wireless networks: Possibilities and fundamental limitations based on available wireless network measurements," *IEEE Signal Process. Mag.*, vol. 22, no. 4, pp. 41–53, Jul. 2005.
[45] N. M. Freris, H. Kowshik, and P. R. Kumar, "Fundamentals of large sensor networks: Connectivity, capacity, clocks, computation," *Proc. IEEE*, vol. 98, no. 11, pp. 1828–1846, Nov. 2010.
[46] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.
[47] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Statist. Soc. Ser. B (Methodol.)*, vol. 58, no. 1, pp. 267–288, 1996.

Georg Kail (M'14) received the B.Sc. and Diplom-Ingenieur (M.Sc.) degrees in electrical engineering/telecommunications and the Dr. techn. (Ph.D.) degree in signal processing from Vienna University of Technology, Vienna, Austria in 2005, 2008, and 2012, respectively. During his master and doctoral studies, he visited UTB Zlín, Czech Republic, ETH Zürich, Switzerland, and ENSEEIHT, Toulouse, France, as a short-term guest researcher. From 2008 to 2013, he was with the Institute of Telecommunications, Vienna University of Technology. During 2013/14, he spent eight months as a Postdoctoral Researcher with the Telecommunications Circuits Laboratory, EPFL Lausanne, Switzerland. Subsequently, as a recipient of an Erwin Schrödinger Fellowship, he was with the Circuits and Systems Group, Delft University of Technology, Delft, The Netherlands for one year. In 2015, he returned to the Institute of Telecommunications in Vienna as a Postdoctoral Research Assistant. His research interests include statistical signal processing with a focus on Bayesian methods and their application to localization tasks.

Sundeep Prabhakar Chepuri (S'11) was born in Bangalore, India in 1986. He received the Bachelors in Engineering degree (*with distinction*) from the PES Institute of Technology, Bangalore, India, in 2007, and the Master of Science degree (*cum laude*) from the Delft University of Technology, The Netherlands, in 2011. He is currently pursuing his Ph.D. at the Faculty of Electrical Engineering, Mathematics and Computer Science of the Delft University of Technology, the Netherlands. He has held positions at Robert Bosch, India, during 2007–2009, and Holst Centre/imec-nl, The Netherlands, during 2010–2011. He has received the "Best Student Paper Award" for his publication at ICASSP 2015 conference in Australia. His general research interest lies in the field of mathematical signal processing, statistical inference, sensor networks, and wireless communications.

Geert Leus (M'01–SM'05–F'12) received the Electrical Engineering degree and the Ph.D. degree in applied sciences from the Katholieke Universiteit Leuven, Belgium, in 1996 and 2000, respectively. Currently, he is an "Antoni van Leeuwenhoek" Full Professor at the Faculty of Electrical Engineering, Mathematics and Computer Science of the Delft University of Technology, Delft, The Netherlands. His research interests are in the area of signal processing for communications.

Prof. Leus received a 2002 IEEE Signal Processing Society Young Author Best Paper Award and a 2005 IEEE Signal Processing Society Best Paper Award. He was the Chair of the IEEE Signal Processing for Communications and Networking Technical Committee, and an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and the IEEE SIGNAL PROCESSING LETTERS. Currently, he is a Member-at-Large to the Board of Governors of the IEEE Signal Processing Society and a member of the IEEE Sensor Array and Multichannel Technical Committee. He finally serves as the Editor in Chief of the *EURASIP Journal on Advances in Signal Processing*.