Multichannel sparse recovery of complex-valued signals using Huber's criterion

Esa Ollila
Department of Signal Processing and Acoustics, Aalto University
P.O.Box 13000, FI-00076 Aalto, Finland

***Abstract*—In this paper, we generalize Huber's criterion to multichannel sparse recovery problem of complex-valued measurements where the objective is to find good recovery of jointly sparse unknown signal vectors from the given multiple measurement vectors which are different linear combinations of the same known elementary vectors. This requires careful characterization of robust complex-valued loss functions as well as Huber's criterion function for the multivariate sparse regression problem. We devise a greedy algorithm based on simultaneous normalized iterative hard thresholding (SNIHT) algorithm. Unlike the conventional SNIHT method, our algorithm, referred to as HUB-SNIHT, is robust under heavy-tailed non-Gaussian noise conditions, yet has a negligible performance loss compared to SNIHT under Gaussian noise. Usefulness of the method is illustrated in source localization application with sensor arrays.**

I. Introduction

In the *multiple measurement vector (MMV) model*, a single measurement matrix is utilized to obtain multiple measurement vectors, i.e., $\mathbf{y}_i = \boldsymbol{\Phi}\mathbf{x}_i + \mathbf{e}_i$, $i = 1, \ldots, Q$ where $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1 \ \cdots \ \boldsymbol{\phi}_N) = \big(\boldsymbol{\phi}_{(1)} \ \cdots \ \boldsymbol{\phi}_{(M)}\big)^{\mathrm{H}}$ is an $M \times N$ *measurement matrix* and $\mathbf{e}_i$ are the (unobserved) random *noise* vectors. Typically there are more column vectors $\boldsymbol{\phi}_i$ than row vectors $\boldsymbol{\phi}_{(j)}$, i.e., $M < N$. The unknown *signal vectors* $\mathbf{x}_i$, $i = 1, \ldots, Q$ are assumed to be *sparse*, i.e., most of the elements are zero. In matrix form, the MMV model is

$$\mathbf{Y} = \boldsymbol{\Phi}\mathbf{X} + \mathbf{E}, \tag{1}$$

where $\mathbf{Y} = (\mathbf{y}_1 \cdots \mathbf{y}_Q) \in \mathbb{C}^{M\times Q}$, $\mathbf{X} = (\mathbf{x}_1 \ \cdots \ \mathbf{x}_Q) \in \mathbb{C}^{N\times Q}$ and $\mathbf{E} = (\mathbf{e}_1 \ \cdots \ \mathbf{e}_Q) \in \mathbb{C}^{M\times Q}$ collect the measurement, the signal and the error vectors, respectively. When $Q = 1$, the model reduces to standard *compressed sensing (CS) model* [1]. The key assumption of MMV model is that the signal matrix $\mathbf{X}$ is K-rowsparse, i.e., at most K rows of $\mathbf{X}$ contain non-zero entries. The *row-support* of $\mathbf{X}$ is the index set of rows containing non-zero elements, $\mathrm{supp}(\mathbf{X}) = \{i \in \{1, \ldots, N\} : x_{ij} \neq 0 \text{ for some } j\}$. When $\mathbf{X}$ is K-rowsparse, i.e., $|\mathrm{supp}(\mathbf{X})| \leq K$, joint estimation can lead both to computational advantages and increased reconstruction accuracy; See [2], [3], [4], [5], [1], [6].

The objective of *multichannel sparse recovery* problem is on finding a row sparse approximation of the signal matrix $\mathbf{X}$ based on knowledge of $\mathbf{Y}$, the measurement matrix $\boldsymbol{\Phi}$ and the sparsity level K. Such a problems arises in electroencephalography and magnetoencephalography (EEG/MEG) [1] blind source separation [7], and direction-of-arrival (DOA) estimation of sources in array and radar processing [8], [9], [10]. Many greedy pursuit CS reconstruction algorithms have been extended for solving MMV problems. These methods, such as simultaneous normalized iterative hard thresholding (SNIHT) algorithm [6] are guaranteed to perform very well provided that suitable conditions (e.g., incoherence of $\boldsymbol{\Phi}$ and non impulsive noise conditions) are met. The derived (worst case) recovery bounds depend linearly on $\|\mathbf{E}\|_2$, so the methods are not guaranteed to provide accurate reconstruction/approximation under heavy-tailed non-Gaussian noise.

In this paper, we generalize Huber's criterion [11, cf. Section 7.7, 7.8] (often referred to as "Huber's approach 2") originally developed for overdetermined linear regression ($M > N$, $Q = 1$) model to the *complex-valued* case and for the more general multivariate *sparse regression* problem. This requires generalizing robust M-estimates of regression (and loss functions) for complex-valued case. In Huber's devise, one estimates the signal matrix and scale of the error terms simultaneously. This is necessary since most robust loss-functions require an estimate of the scale. Using Huber's criterion in the MMV model one may elegantly estimate both the sparse signal matrix and the scale of the errors simultaneously. In particularly, we are able to circumvent the problem of obtaining a preliminary robust scale estimate which is a challenging problem in ill-posed multivariate sparse regression model since the support of $\mathbf{X}$ and hence the contributing elementary vectors of $\boldsymbol{\Phi}$ on measurements are not known. In earlier related work Huber's approach 2 has been considered for Lasso-type real-valued linear regression setting in [12], [13] and real-valued compressed sensing in [14]. For our multichannel sparse recovery problem, we devise SNIHT algorithm which results in a simple, computationally efficient and scalable approach for solving the MMV sparse reconstruction problem.

Let us offer a brief outline of the paper. In Section II, we give necessary notations and definitions as well as provide motivation and background of robust sparse recovery problem. Robust complex-valued loss functions and their properties are outlined in Section III and a generalization of Huber's loss function for complex measurements is given. Then, in Section IV we formulate Huber's criterion for MMV model and the related SNIHT algorithm, called HUB-SNIHT, is derived in Section V. Finally, we illustrate the usefullness of the method in source localization application in Section VI.

II. Background

A. Notations

For a matrix $\mathbf{A} \in \mathbb{C}^{M\times N}$ and an index set Γ of cardinality $|\Gamma| = K$, we denote by $\mathbf{A}_\Gamma$ (resp. $\mathbf{A}_{(\Gamma)}$) the $M \times K$ (resp. $K \times N$) matrix restricted to the columns (resp. rows) of $\mathbf{A}$

978-1-4799-7420-7/15/$31.00 ©2015 IEEE

indexed by the set Γ. The ith column vector of $\mathbf{A}$ is denoted by $\mathbf{a}_i$ and the hermitian transpose of the ith row vector of $\mathbf{A}$ by $\mathbf{a}_{(i)}$, $\mathbf{A} = (\mathbf{a}_1 \ \cdots \ \mathbf{a}_N) = (\mathbf{a}_{(1)} \ \cdots \ \mathbf{a}_{(M)})^{\mathrm{H}}$. Furthermore, if $f : \mathbb{C} \to \mathbb{C}$, then $f(\mathbf{A})$ refers to element-wise application of the function to its matrix valued argument, so $f(\mathbf{A}) \in \mathbb{C}^{M\times N}$ with $[f(\mathbf{A})]_{ij} = f(a_{ij})$.

The usual Euclidean norm on vectors will be written as $\|\cdot\|$. The matrix space $\mathbb{C}^{M\times N}$ is equipped with the usual Hermitian inner product

$$\langle \mathbf{A}, \mathbf{B}\rangle = \mathrm{Tr}(\mathbf{B}^{\mathrm{H}}\mathbf{A}) = \sum_{i=1}^{M}\sum_{j=1}^{N} a_{ij}b_{ij}^{*}$$

where the trace of a (square) matrix is the sum of diagonal entries. We define the weighted inner product as

$$\langle \mathbf{A}, \mathbf{B}\rangle_{\mathbf{W}} = \sum_{i=1}^{M}\sum_{j=1}^{N} w_{ij}a_{ij}b_{ij}^{*}$$

where $\mathbf{W}$ is $M \times N$ real matrix of positive weights. Note that $\langle \mathbf{A}, \mathbf{B}\rangle_{\mathbf{W}}$ reduces to conventional inner product when $\mathbf{W}$ is a matrix of ones. The Frobenius norm is given by the inner product as $\|\mathbf{A}\| = \sqrt{\langle \mathbf{A}, \mathbf{A}\rangle}$ and $\|\mathbf{A}\|_{\mathbf{W}} = \sqrt{\langle \mathbf{A}, \mathbf{A}\rangle_{\mathbf{W}}}$ denotes the weighted Frobenius norm. The row-ℓ_0 quasi-norm of $\mathbf{A}$ is the number of nonzero rows, i.e., $\|\mathbf{A}\|_0 = |\operatorname{supp}(\mathbf{A})|$. Hence the assumption that the signal matrix $\mathbf{X} \in \mathbb{C}^{N\times Q}$ is K-rowsparse in the MMV model is equivalent with the statement that $\|\mathbf{X}\|_0 \leq K$.

We use $H_K(\cdot)$ to denote the *hard thresholding operator*: for a matrix $\mathbf{X} \in \mathbb{C}^{N\times Q}$, $H_K(\mathbf{X})$ retains the elements of the K rows of $\mathbf{X}$ that possess largest ℓ_2-norms and set elements of the other rows to zero. Notation $\mathbf{X}|_{\Gamma}$ refers to sparsified version of $\mathbf{X}$ such that the entries in the rows indexed by set Γ remain unchanged while all other rows have all entries set to 0.

B. Robust constrained optimization problem

Suppose that the error terms e_{ij} are i.i.d. continuous random variables from a circular distribution [15] with p.d.f. $f(e) = (1/\sigma)f_0(e/\sigma)$, where $f_0(e)$ denotes the standard form of the density and $\sigma > 0$ is the scale parameter. If the scale is known, then a reasonable approach for solving the simultaneous sparse recovery problem is to minimize a distance criterion of residuals,

$$D_{\rho}\left(\frac{\mathbf{Y}-\mathbf{\Phi X}}{\sigma}\right) = \sum_{i=1}^{M}\sum_{j=1}^{Q}\rho\left(\frac{y_{ij}-\boldsymbol{\phi}_{(i)}^{\mathrm{H}}\mathbf{x}_j}{\sigma}\right) \qquad (2)$$

for some suitable loss function $\rho(\cdot)$ subject to K-rowsparsity constraint $\|\mathbf{X}\|_0 \leq K$. For conventional least squares (LS) loss function $\rho(e) = |e|^2$, the scale can be factored out from the objective function, and the minimization problem reduces to

$$\min_{\mathbf{X}} \|\mathbf{Y}-\mathbf{\Phi X}\|^2 \quad \text{subject to} \quad \|\mathbf{X}\|_0 \leq K.$$

The well-known problem with LS minimization is that it gives a very small weight on small residuals and a strong weight on large residuals, implying that even a single large outlier can have a large influence on the obtained result.

At least two problems arises when using conventional robust loss functions in (2). First, commonly used robust loss functions in robust statistics such as Huber's or Tukey's loss functions require an estimate of scale σ. Obtaining a reliable robust estimate of scale is a difficult problem. It involves obtaining a K-rowsparse robust preliminary estimate $\hat{\mathbf{X}}_0$ of the signal matrix and then computing robust scale estimate based on the resulting residual matrix $\mathbf{R}_0 = \mathbf{Y} - \mathbf{\Phi}\hat{\mathbf{X}}_0$. Second problem is that robust loss functions are defined in the real-valued case and some thought must be given on special properties of complex-valued loss functions. These problems are addressed next in Section III and Section IV.

III. Loss functions: complex valued case

We start by giving a proper definition of a loss function ρ.

Definition 1: Function $\rho : \mathbb{C} \to \mathbb{R}_0^+$ is called a *loss function* if it verifies:

(L1) ρ is circularly symmetric, $\rho(e^{\jmath\theta}x) = \rho(x)$, $\forall\theta \in \mathbb{R}$.

(L2) $\rho(0) = 0$. Furthermore, ρ is $\mathbb{R}$-differentiable function and increasing in $|e| > 0$.

Let us first note that condition (L1) is equivalent with the statement

$$\rho(x) = \rho_0(|x|) \qquad (3)$$

for some $\rho_0 : \mathbb{R}_0^+ \to \mathbb{R}_0^+$. The fact that (3) $\Rightarrow$ (L1) is obvious and the converse can be derived by invariance arguments. This illustrates that ρ is not $\mathbb{C}$-differentiable (i.e., holomorphic or analytic function). This is of course natural since only functions that are *both* holomorphic *and* real-valued are constants. The complex derivative of ρ w.r.t. $x^* = (x_R + \jmath x_I)^*$ is

$$\psi(x) = \frac{\partial}{\partial x^*}\rho(x) = \frac{1}{2}\left(\frac{\partial\rho}{\partial x_R} + \jmath\frac{\partial\rho}{\partial x_I}\right)$$

which will be referred in the sequel as the *score function*. Since $\rho(e) = \rho_0(|e|)$, we can write ψ using basic rules of complex differentiation [16] in the form

$$\psi(x) = \frac{1}{2}\rho_0'(|x|)\mathrm{sign}(x),$$

where

$$\mathrm{sign}(e) = \begin{cases} e/|e|, & \text{for } e \neq 0 \\ 0, & \text{for } e = 0 \end{cases}$$

is the complex *signum function* and ρ_0' denotes the real derivative of the real-valued function ρ_0. In order to make minimization of (2) possible by simple gradient descent type algorithms, we narrow down the set of loss functions by imposing the assumption:

(L3) $\rho : \mathbb{C} \to \mathbb{R}_0^+$ is a convex function

For example, the conventional LS loss function $\rho(x) = |x|^2$ verifies assumptions (L1)-(L3). In this case, $\rho_0(r) = r^2$ and the score function is $\psi(x) = x$. In this paper, we assume that the loss function verifies (L1)-(L3).

We define *Huber's loss function* in the complex case as

$$\rho_{H,c}(e) = \begin{cases} |e|^2, & \text{for } |e| \leq c \\ 2c|e| - c^2, & \text{for } |e| > c, \end{cases} \qquad (4)$$

where c is a user-defined *threshold* that influences the degree of robustness and efficiency of the method. Huber's function is a hybrid of ℓ_2 and ℓ_1 loss functions, using ℓ_2-loss for relatively small errors and ℓ_1-loss for relatively large errors. It verifies conditions (L1)-(L3). Huber's score (ψ-)function is

$$\psi_{H,c}(e) = \begin{cases} e, & \text{for } |e| \le c \\ c\,\mathrm{sign}(e), & \text{for } |e| > c \end{cases}$$

Note that Huber's ψ is a winsorizing (clipping) funtion: the smaller the c, the more clipping is actioned on the residuals.

IV. Huber's criterion for multichannel sparse recovery

As discussed earlier, the scale σ of the error terms is unknown and needs to be estimated jointly with the signal matrix. We discuss here how this can be done elegantly using Huber's approach 2. First note that Maximum likelihood (ML-)approach for solving the unknown $\mathbf{X}$ and σ leads to minimizing the negative log-likelihood function of the form

$$Q_{ML}(\mathbf{X},\sigma) = (MQ)\log\sigma + \sum_{i=1}^{M}\sum_{j=1}^{Q}\rho\left(\frac{y_{ij} - \boldsymbol{\phi}_{(i)}^{\mathrm{H}}\mathbf{x}_i}{\sigma}\right)$$

where $\rho(e) = -\log f_0(e)$ depends on the underlying standard form of the density $f_0(e)$ of the error terms. Then, one could replace the ML loss function ρ with a robust loss function which need not be related to any circular density $f_0(\cdot)$, e.g., the Huber's loss function. The negative log-likelihood function is however not convex in $(\mathbf{X},\sigma)$. This follows since $Q_{ML}(\mathbf{X},\sigma)$ is not convex in σ (for fixed $\mathbf{X}$) and hence cannot be jointly convex.

Huber [11] proposed an elegant devise to circumvent the above problem. See also [12] for further study of Huber's approach. We generalize the Huber's approach 2 for the complex multivariate regression case and minimize

$$Q(\mathbf{X},\sigma) = \alpha(MQ)\sigma + \sum_{i=1}^{M}\sum_{j=1}^{Q}\rho\left(\frac{y_{ij} - \boldsymbol{\phi}_{(i)}^{\mathrm{H}}\mathbf{x}_i}{\sigma}\right)\sigma, \quad (5)$$

where $\alpha > 0$ is a fixed *scaling factor*. Important feature of the objective function is that it is jointly convex in $(\mathbf{X},\sigma)$ given that ρ is convex. In addition the minimizer $\hat{\mathbf{X}}$ preserves the same theoretical robustness properties (such as bounded influence function) as the minimizer in the model where σ is assumed to be known (fixed). This is not the case for the ML-objective function $Q_{ML}(\mathbf{X},\sigma)$.

The stationary point of (5) can be found by setting the complex matrix derivative of Q w.r.t. $\mathbf{X}^*$ and the real derivative of Q w.r.t. σ to zero. Simple calculations then show that the minimizer $(\hat{\mathbf{X}},\hat{\sigma})$ is a solution to a pair of M-estimating equations:

$$\boldsymbol{\Phi}^{\mathrm{H}}\psi\left(\frac{\mathbf{R}}{\sigma}\right) = \mathbf{0} \quad (6)$$

$$\frac{1}{MQ}\sum_{i=1}^{M}\sum_{j=1}^{Q}\chi\left(\frac{y_{ij} - \boldsymbol{\phi}_{(i)}^{\mathrm{H}}\mathbf{x}_j}{\sigma}\right) = \alpha \quad (7)$$

where $\mathbf{R} = \mathbf{Y} - \boldsymbol{\Phi}\mathbf{X}$ and $\chi : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ is defined as

$$\chi(t) = \rho_0'(t)t - \rho_0(t). \quad (8)$$

Recall that notation $\psi(\mathbf{R})$ refers to element-wise application of ψ-function to its matrix valued argument, so $[\psi(\mathbf{R})]_{ij} = \psi(r_{ij})$. Thus if ρ is convex and the MMV model is overdetermined with non-sparse $\mathbf{X}$, solving the above M-estimating equations would give the global minimum of (5).

The scaling factor α in (5) is chosen so that the obtained scale estimate $\hat{\sigma}$ is Fisher-consistent for the unknown scale σ when $e_{ij} \sim \mathbb{C}\mathcal{N}(0,\sigma^2)$, which due to (7) is chosen so that

$$\alpha = \mathbb{E}[\chi(e)], \quad e \sim \mathbb{C}\mathcal{N}(0,1).$$

For many loss functions, α can be computed in closed-form. For example, for Huber's function (4) the χ-function in (8) becomes

$$\chi_{H,c}(e) = |\psi_{H,c}(e)|^2 = \begin{cases} |e|^2, & \text{for } |e| \le c \\ c^2, & \text{for } |e| > c, \end{cases}$$

and the concistency factor $\alpha = \alpha(c)$ can be easily solved in closed-form by elementary calculus as

$$\alpha = c^2(1 - F_{\chi_2^2}(2c^2)) + F_{\chi_4^2}(2c^2). \quad (9)$$

Note that α depends on the threshold c. We will choose threshold c as $c^2 = (1/2)F^{-1}_{\chi_2^2}(q)$ for $q \in (0,1)$. The rationale behind this choice is that under Gaussian errors, $2|e|^2/\sigma^2 \sim \chi_2^2$. Hence a sensible choice is to determine c so that $2c^2$ is the qth upper quantile of the χ_2^2-distribution. The choice $q \to 1$, implies $c^2 \to \infty$ and hence no-trimming of the residuals. In our simulations we use $q = 0.8$ which yields $c = 1.269$. The smaller the c (and hence q) the more trimming is actioned on residuals.

V. SNIHT algorithm for Huber's criterion

Our aim is at solving

$$\min_{\mathbf{X},\sigma} Q(\mathbf{X},\sigma) \text{ subject to } \|\mathbf{X}\|_0 \le K.$$

This problem is combinatorial (i.e., NP-hard) but *greedy pursuit approaches* can be devised. Thus due to biconvexity of the objective function, we can use Huber's loss function $\rho_{H,c}(e)$ and greedy pursuit NIHT algorithm can be devised to compute an approximate solution. Recall that NIHT is a *projected gradient descent* method that is known to offer efficient and scalable solution for K-sparse approximation problem [17]. NIHT updates the estimate of $\mathbf{X}$ by taking steps towards the direction of the negative gradient followed by projection onto the constrained space.

In Huber's criterion, if we consider σ fixed at a value $\sigma = \sigma^{n+1}$ (the value of σ at $(n+1)$th iteration), the simultaneous NIHT (SNIHT) update of the signal matrix becomes

$$\mathbf{X}^{n+1} = H_K\big(\mathbf{X}^n + \mu^{n+1}\boldsymbol{\Phi}^{\mathrm{H}}\mathbf{R}_\psi^n\big)$$

where μ^{n+1} is the update of the stepsize at $(n+1)$th iteration and

$$\mathbf{R}_\psi^n = \psi\left(\frac{\mathbf{R}^n}{\sigma^{n+1}}\right)\sigma^{n+1}$$

will be referred to as *pseudo-residual*. Note that $-\nabla_{\mathbf{X}^*}\rho\big(\frac{\mathbf{Y}-\boldsymbol{\Phi}\mathbf{X}}{\sigma^{n+1}}\big)\,(\sigma^{n+1})^2 = \boldsymbol{\Phi}^{\mathrm{H}}\mathbf{R}_\psi^n$. The scale is updated

(consider signal matrix $\mathbf{X}$ fixed at a value $\mathbf{X} = \mathbf{X}^n$) using (7) by a fixed-point iteration

$$(\sigma^{n+1})^2 = \frac{(\sigma^n)^2}{\alpha}\frac{1}{MQ}\sum_{i=1}^{M}\sum_{j=1}^{Q}\chi\left(\frac{r_{ij}^n}{\sigma^n}\right),$$

where $\mathbf{R}^n = \mathbf{Y} - \mathbf{\Phi}\mathbf{X}^n$

The pseudo-code for the SNIHT algorithm in the case that the loss function ρ is Huber's function (4) is given in Algorithm 1. We refer to this algorithm as HUB-SNIHT in the sequel. The steps 3-9 can be divided to 3 stages described below: *scale stage* (Steps 3, 4) build up the scale update σ^{n+1}, *signal stage* (Steps 5, 7, 8, 9) build up the K-sparse signal update $\mathbf{X}^{n+1}$ and the support Γ^{n+1}, and *stepsize stage* (Step 7) computes the optimal stepsize update for the gradient descent move. The computation of the stepsize will be described in the next two paragraphs. Note that it is possible to tune the algorithm for different applications by simply altering the criterion for halting the algorithm. Matlab function is available at http://users.spa.aalto.fi/esollila/software.html.

Algorithm 1: HUB-SNIHT algorithm

input : $\mathbf{Y}$, $\mathbf{\Phi}$, sparsity K, trimming threshold c.
output : $(\mathbf{X}^{n+1}, \sigma^{n+1}, \Gamma^{n+1})$ estimates of $\mathbf{X}$, σ and $\Gamma = \mathrm{supp}(\mathbf{X})$.
initialize: $\mathbf{X}^0 = \mathbf{0}$, $\mu^0 = 0$, $n = 0$, $\Gamma^0 = \emptyset$, $\alpha = \alpha(c)$.

1 $\sigma^0 = 1.201 \cdot \mathrm{median}(|y_{ij}|, i = 1, \ldots, M, j = 1, \ldots, Q)$
2 $\Gamma^0 = \mathrm{supp}(H_K(\mathbf{\Phi}^{\mathrm{H}}\psi_{H,c}(\mathbf{Y}/\sigma^0)))$
while *halting criterion false* **do**
3 $\quad \mathbf{R}^n = \mathbf{Y} - \mathbf{\Phi}\mathbf{X}^n$
4 $\quad (\sigma^{n+1})^2 = \frac{(\sigma^n)^2}{\alpha}\frac{1}{MQ}\sum_{i=1}^{M}\sum_{j=1}^{Q}\left|\psi_{H,c}\left(\frac{r_{ij}^n}{\sigma^n}\right)\right|^2$
5 $\quad \mathbf{R}_\psi^n = \psi_{H,c}\left(\frac{\mathbf{R}^n}{\sigma^{n+1}}\right)\sigma^{n+1}$
6 $\quad \mathbf{G}^n = \mathbf{\Phi}^{\mathrm{H}}\mathbf{R}_\psi^n$
7 $\quad \mu^{n+1} = \texttt{CompStepsize}(\mathbf{R}^n, \mathbf{\Phi}, \mathbf{G}, \Gamma^n, \mu^n, \sigma^{n+1})$
8 $\quad \mathbf{X}^{n+1} = H_K(\mathbf{X}^n + \mu^{n+1}\mathbf{G}^n)$
9 $\quad \Gamma^{n+1} = \mathrm{supp}(\mathbf{X}^{n+1})$
10 $\quad n = n + 1$
end

As was noted in [17], stepsize selection is very important for convergence and needs to be adaptively controlled at each iteration. Given the found support Γ^n is correct, we choose μ^{n+1} as the minimizer of the convex objective function (2) for fixed scale at σ^{n+1} in the gradient ascent direction $\mathbf{X}^n + \mu\mathbf{G}^n|_{\Gamma^n}$, i.e.

$$\begin{aligned} L(\mu) &= D_{\rho_{H,c}}\left(\frac{\mathbf{Y} - \mathbf{\Phi}\left(\mathbf{X}^n + \mu\mathbf{G}^n|_{\Gamma^n}\right)}{\sigma^{n+1}}\right) \\ &= D_{\rho_{H,c}}\left(\frac{\mathbf{R}^n - \mu\mathbf{B}^n}{\sigma^{n+1}}\right) \end{aligned} \tag{10}$$

where $\mathbf{R}^n = \mathbf{Y} - \mathbf{\Phi}\mathbf{X}^n$ and $\mathbf{B}^n = \mathbf{\Phi}_{\Gamma^n}\mathbf{G}^n_{(\Gamma^n)}$. This reduces to minimizing a simple linear regression (M-)estimation problem where the response is $\mathbf{r} = \mathrm{vec}(\mathbf{R}^n)$ and the predictor is $\mathbf{b} = \mathrm{vec}(\mathbf{B}^n)$. It is easy to show (details omitted) that the minimizer $\hat{\mu}$ of $L(\mu)$ is the unique solution to a fixed point (FP) equation $\mu = H(\mu)$, where

$$H(\mu) = \|\mathbf{B}^n\|^{-2}_{\mathbf{W}(\mu)}\,\mathrm{Re}(\langle\mathbf{R}^n, \mathbf{B}^n\rangle_{\mathbf{W}(\mu)}) \tag{11}$$

where the right hand side depends on the unknown μ via the weight matrix $\mathbf{W}(\mu)$, defined as

$$\mathbf{W}(\mu) = w_{H,c}\left(\frac{\mathbf{R}^n - \mu\mathbf{B}^n}{\sigma^{n+1}}\right),$$

where $w_{H,c}$ is a weight function based on Huber's loss function, defined as

$$w_{H,c}(e) = \frac{\psi_{H,c}(e)}{e} = \begin{cases} 1, & \text{for } |e| \le c \\ c/|e|, & \text{for } |e| > c \end{cases}.$$

If the loss function is LS-loss $\rho(e) = |e|^2$ (equivalent to Huber's function when $c \to \infty$), then the minimizer of (10) is easily found in closed form since in this case $\mathbf{W}(\mu)$ is equal to a matrix of ones. Hence the FP equation is explicit and the solution is $\mu^{n+1} = \|\mathbf{G}^n_{(\Gamma^n)}\|^2/\|\mathbf{\Phi}_{\Gamma^n}\mathbf{G}^n_{(\Gamma^n)}\|^2$. This is indeed the same stepsize used in conventional SNIHT [6].

For Huber's loss function, the minimizer of (10) can be found by running the FP iterations until convergence (with initial value $\mu_0 > 0$). Instead, we use approximate of the solution given by 1-step FP iterate with initial value given by the previous stepsize μ^n. In other words, in Step 7, the update μ^{n+1} is computed as $\mu^{n+1} = H(\mu^n)$.

VI. Application to Source Localization

We consider sensor array consisting of M sensors that receives K narrowband incoherent farfield plane-wave sources from a point source ($M > K$). At discrete time t, the *array output* (snapshot) $\mathbf{y}(t) \in \mathbb{C}^M$ is a weighted linear combination of the signal waveforms $\mathbf{x}(t) = (x_1(t), \ldots, x_K(t))^\top$ corrupted by additive noise $\mathbf{e}(t) \in \mathbb{C}^M$, $\mathbf{y}(t) = \mathbf{A}(\boldsymbol{\theta})\mathbf{x}(t) + \mathbf{e}(t)$, where $\mathbf{A} = \mathbf{A}(\boldsymbol{\theta})$ is the $M \times K$ *steering matrix* parametrized by the vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^\top$ of (distinct) unknown direction-of-arrivals (DOA's) of the sources. Each column vector $\mathbf{a}(\theta_i)$, called the *steering vector*, represents a point in known array manifold $\mathbf{a}(\theta)$. The objective of sensor array source localization is to find the DOA's of the sources, i.e., to identify the steering matrix $\mathbf{A}(\boldsymbol{\theta})$ parametrized by $\boldsymbol{\theta}$. We assume that the number of sources K is known.

As in [8], we cast the source localization problem as a multichannel sparse recovery problem. We construct an overcomplete $M \times N$ steering matrix $\mathbf{A}(\tilde{\boldsymbol{\theta}})$, where $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_N)^\top$ represents a sampling grid of all source locations of interest. If $\tilde{\boldsymbol{\theta}}$ contains the true DOA's θ_i, $i = 1, \ldots, K$, then the measurement matrix $\mathbf{Y} = (\mathbf{y}(t_1) \quad \cdots \quad \mathbf{y}(t_Q)) \in \mathbb{C}^{M\times Q}$ consisting of snapshots at time instants $t_1, \ldots, t_Q$ can be *exactly* modelled as MMV model (1), where the signal matrix $\mathbf{X} \in \mathbb{C}^{N\times Q}$ is K-rowsparse matrix with source signal sequences as its non-zero row vectors. Thus identifying the source locations is equivalent to identifying the support $\Gamma = \mathrm{supp}(\mathbf{X})$ since any $i \in \Gamma$ maps to a DOA $\tilde{\theta}_i$ in the grid. Since the steering matrix $\mathbf{A}(\tilde{\boldsymbol{\theta}})$ is completely known, we can use HUB-SNIHT method to identify the support.

We assume that $K = 2$ independent (spatially and temporally) complex circular Gaussian source signals of equal

power σ_x^2 arrive on an uniform linear array (ULA) of $M = 20$ sensors with half a wavelength inter-element spacing from DOA's $\theta_1 = 0^o$ and $\theta_2 = 8^o$. In this case, the array manifold is $\mathbf{a}(\theta) = (1, e^{-\jmath\pi \sin(\theta)}, \cdots, e^{-\jmath\pi(M-1)\sin(\theta)})^\top$. The noise matrix $\mathbf{E} \in \mathbb{C}^{M\times Q}$ has i.i.d. elements following inverse Gaussian compound Gaussian (IG-CG) distribution [18] with shape parameter $\lambda = 0.1$ and unit variance. CG-IG distribution is heavy-tailed and has been shown to accurately model radar clutter in [18]. Note that the covariance matrix of the snapshot is $\mathrm{Cov}(\mathbf{y}(t_i)) = \sigma_x^2 \mathbf{A}(\boldsymbol{\theta})\mathbf{A}(\boldsymbol{\theta})^{\mathrm{H}} + \mathbf{I}_M$, so we may use the popular MUSIC method to localize the sources. In other words, we search for $K = 2$ peaks of the MUSIC pseudospectrum in the grid. We use a uniform grid $\tilde{\boldsymbol{\theta}}$ on $[-90, 90]$ with 2^o degree spacing, thus containing the true DOA's. For the source localization application, we make the following modifcation to the algorithm: In Step 1 of HUB-SNIHT algorithm, we locate the K largest peaks of rownorms of $\boldsymbol{\Phi}^{\mathrm{H}}\psi_{H,c}(\mathbf{Y})$ instead of taking Γ^0 as indices of K largest rownorms of $\boldsymbol{\Phi}^{\mathrm{H}}\psi_{H,c}(\mathbf{Y})$.

We then use SNIHT, HUB-SNIHT and MUSIC to identify the support (which gives the DOA estimates) and compute the empirical *probability of exact recovery* (PER) rates and the relative frequency of DOA estimates in the grid based on 1000 MC runs. Full PER rate $= 1$ implies that the support Γ (and hence DOA's) were correctly identified in all MC trials. Such a case is shown in upper plot of Figure 1 for HUB-SNIHT when the number of snapshots is $Q = 50$ and the SNR is -10 dB. The PER rate of HUB-SNIHT was 0.99, but PER rates of SNIHT and MUSIC were considerably lower, 0.81 and 0.94, respectively. In the second setting, we lower the SNR to -20 dB. In this case, the conventional SNIHT and MUSIC methods fail completely and provide nearly a uniform frequency on the grid. This is illustrated in the middle plot of Figure 1. Note that the robust HUB-SNIHT provides high peaks on the correct DOA's. The PER rates of SNIHT, HUB-SNIHT and MUSIC were $0.02, 0.48$ and 0.01, respectively. Thus only HUB-SNIHT is able to offer good localization of the sources whereas the non-robust methods do not provide much better performance than a random guess. In the 3rd setting, we alter the set-up of 1st setting by decreasing the number of snapshots from $Q = 50$ as low as $Q = 5$. The performance differences between the methods are now more significant as is illustrated in the lower plot of Figure 1. In this case the PER rates of SNIHT, HUB-SNIHT and MUSIC were 0.19, 0.57 and 0.37, respectively. Again, the HUB-SNIHT performed the best.

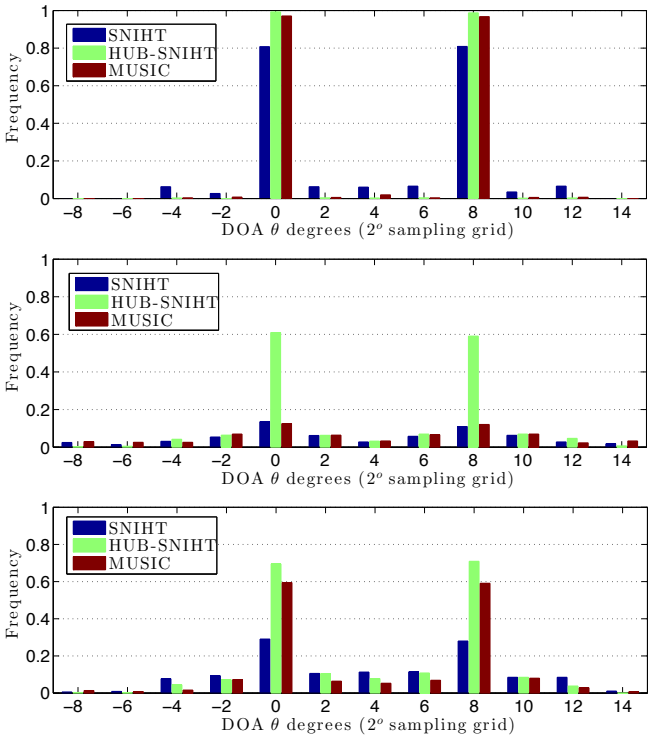

Fig. 1. Bar plots of relative frequency of DOA estimates. Two equal power independent Gaussian sources arrive from DOA 0^o and 8^o and the noise has i.i.d. elements from IG-CG distribution with unit variance and shape $\lambda = 0.1$. SNR $= -10$ dB and $Q = 50$ (upper plot), SNR $= -20$ dB and $Q = 50$ (middle plot) and SNR $= -10$ dB and $Q = 5$ (lower plot).

REFERENCES

[1] M. F. Duarte and Y. C. Eldar, "Structured compressed sensing: From theory to applications," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4053–4085, 2011.

[2] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. part I: Greedy pursuit," *Signal Processing*, vol. 86, pp. 572–588, 2006.

[3] J. A. Tropp, "Algorithms for simultaneous sparse approximation. part II: Convex relaxation," *Signal Processing*, vol. 86, pp. 589–602, 2006.

[4] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors," *IEEE Trans. Signal Process.*, vol. 54, no. 12, pp. 4634–4643, 2006.

[5] Y. C. Eldar and H. Rauhut, "Average case analysis of multichannel sparse recovery using convex relaxation," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 505–519, 2010.

[6] J. D. Blanchard, M. Cermak, D. Hanle, and Y. Jin, "Greedy algorithms for joint sparse recovery," *IEEE Trans. Signal Process.*, vol. 62, no. 7, pp. 1694 – 1704, 2014.

[7] R. Gribonval and M. Zibulevsky, *Handbook of Blind Source Separation*. Oxford, UK: Academic Press, 2010, ch. Sparse component analysis, pp. 367–420.

[8] D. Malioutov, M. Çetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3010–3022, 2005.

[9] Y. Wang, G. Leus, and A. Pandharipande, "Direction estimation using compressive sampling array processing," in *IEEE Workshop on Statistical Signal Processing (SSP'09)*, 2009, pp. 626–629.

[10] S. Fortunati, R. Grasso, F. Gini, M. S. Greco, and K. LePage, "Single-snapshot DOA estimation by using compressed sensing," *EURASIP J. Adv. Signal Process.*, vol. 2014, no. 1, pp. 1–17, 2014.

[11] P. J. Huber, *Robust Statistics*. New York: Wiley, 1981.

[12] A. B. Owen, "A robust hybrid of lasso and ridge regression," *Contemporary Mathematics*, vol. 443, pp. 59–72, 2007.

[13] S. Lambert-Lacroix, L. Zwald *et al.*, "Robust regression through the hubers criterion and adaptive lasso penalty," *Electronic Journal of Statistics*, vol. 5, pp. 1015–1053, 2011.

[14] E. Ollila, H.-J. Kim, and V. Koivunen, "Robust iterative hard thresholding for compressed sensing," in *Proc. IEEE Int'l Symp. Communications, Control, and Signal Processing (ISCCSP'14)*, Athens, Greece, May 21 – 23, 2014, pp. 226–229.

[15] E. Ollila, J. Eriksson, and V. Koivunen, "Complex elliptically symmetric random variables – generation, characterization, and circularity tests," *IEEE Trans. Signal Process.*, vol. 59, no. 1, pp. 58–69, 2011.

[16] J. Eriksson, E. Ollila, and V. Koivunen, "Essential statistics and tools for complex random variables," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5400–5408, 2010.

[17] T. Blumensath and M. E. Davies, "Normalized iterative hard thresholding: guaranteed stability and performance," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 298–309, 2010.

[18] E. Ollila, D. E. Tyler, V. Koivunen, and H. V. Poor, "Compound-gaussian clutter modelling with an inverse gaussian texture distribution," *IEEE Signal Process. Lett.*, vol. 19, no. 12, pp. 876–879, 2012.